IMPROVING PRINCIPAL PRACTICE: ADDRESSING TEACHER EVALUATION THROUGH DATA ANALYSIS AND STRATEGIC PROFESSIONAL DEVELOPMENT

A disquisition submitted to the faculty of the Graduate School of
Western Carolina University in partial fulfillment of the
Requirements for the degree of Doctor of Education

By

Heather Pendry Mullins

Director: Dr. Kathleen Topolka-Jorissen
Associate Professor
Department of Human Services

Committee Members:
Dr. Robert Crow, Human Services
Dr. John Sherlock, Human Services
Dr. Aron Gabriel, Assistant Superintendent, Newton-Conover

February 2016

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACS – Appalachian County Schools
AP – Assistant Principal
CAO – Chief Academic Officer
CMS – Concordia Middle School
CPS – Concordia Public Schools
CWT – Classroom Walkthrough Tool
DLP – Distinguished Leadership in Practice
ELT – Executive Leadership Team
EOC – End of Course
EOG – End of Grade
ESEA – Elementary and Secondary Education Act
ESSA – Every Child Succeeds Act
EVAAS – Educator Value-Added Assessment System
IHE – Institution of Higher Education
LEA – Local Education Agency
MET – Measures of Effective Teaching
MRM – Multivariate model
NCDPI – North Carolina Department of Public Instruction
NCEES – North Carolina Educator Effectiveness System
McREL – Mid-Continent Research for Education and Learning
NCLB – No Child Left Behind
NIC – Networked Improvement Community
OCT – Observation Calibration Training
PCOLT – Professional Community, Organizational Learning, and Trust
PDSA – Plan, Do, Study, Act
PDP – Professional Development Plan
PLC – Professional Learning Community
RESA – Regional Education Service Alliance
RttT – Race to the Top
SAMR – Substitution, Augmentation, Modification, and Redefinition
SAS – Statistical Analysis System
SS – Scoring Study
SS1 – Scoring Study 1
SS2 – Scoring Study 2
SS3 – Scoring Study 3
TPACK – Technological Pedagogical and Content Knowledge
TPAI-R – Teacher Appraisal Instrument – Revised
URM – Univariate model
VAM – Value-Added Model

ABSTRACT

IMPROVING PRINCIPAL PRACTICE: ADDRESSING TEACHER EVALUATION THROUGH DATA ANALYSIS AND STRATEGIC PROFESSIONAL DEVELOPMENT

Heather Pendry Mullins, Ed.D.

Western Carolina University (February 2016)

Director: Dr. Kathleen Topolka-Jorissen

This improvement project was developed to address the discrepancy between how principals rate and evaluate teachers and the value-added measures teachers receive based on student growth data from standardized tests. Specifically, this project engaged a group of principals in a professional development program designed to improve their rater agreement against criterion measures in a commercially developed program. Employing improvement science methods, a multi-tiered process of interventions was designed and implemented with the goals of improving target agreement, decreasing discrepancy, and eliminating scoring bias among the twelve evaluators who participated in the study in the Concordia School District. After each intervention, data were analyzed to determine appropriate subsequent interventions. Through collaboration with leaders from the North Carolina Department of Public Instruction (NCDPI) the project was enhanced and extended through face-to-face opportunities and follow-up sessions.

Qualitative and quantitative data from the project indicate that evaluation ratings improved among the participants. Due to the collaborative nature of the project between a district and the state department of education, this prototype for improving the quality of evaluations has implications to serve as a statewide model.

CHAPTER 1: INTRODUCTION

The concept of educator effectiveness has been one that has stimulated the curiosity and interest of educational scholars for decades. Over 30 years ago Berliner (1982) suggested that educator effectiveness was more than a test score and could only be determined accurately by a "connoisseur of classrooms" to know whether a teacher is truly effective. In an interview with Ron Brandt, Berliner concluded that the definition of effectiveness was elusive, but having instructional leaders who could gauge the climate of the classroom and offer coaching and support to teachers who needed guidance was imperative to improving teacher effectiveness. Thirty years later, scholars had developed a more cohesive definition of educator effectiveness. Darling-Hammond (2006) indicated that deep content and pedagogical knowledge, experience, and successful completion of the demands of a teacher licensure program were the leading factors in identifying effective teachers. In the last ten years, however, one national movement has redefined not only educator effectiveness but also how administrators evaluate teachers.

Since the inception of educator value-added evaluation models (VAMs), opinions about what makes a teacher effective have continued to evolve. With more data at the fingertips of leaders and policymakers across our nation, a new redefinition of educator effectiveness is emerging. Growth has taken the spotlight in education, and it has become the new normal for measuring a teacher's overall effectiveness. Not only have teachers, schools, and districts been either praised or criticized for their growth or lack thereof, but the evaluation data principals submit for teachers have also been the center of much debate as there appears to be a lack of correlation between teachers' value-added

growth measures and how their principals evaluate them. Researchers have reported modest to low correlations between ratings teachers receive from evaluators and value-added growth measures (Bell, Gitomer, McCaffrey, Hambre, Pianta, & Qi, 2012 & Strong, Gargani, & Hacifazlioğlu, 2011). It seems that the keys to improving teacher effectiveness may lie in improving evaluator effectiveness, accuracy, and capacity to provide instructional support. In my experience as an educational leader, the vast majority of teachers I have worked with truly desire to be exceptional teachers. They work long hours and are willing to make changes to meet students' needs. However, when they are not meeting the standard set forth by the school or district, they need and deserve high-quality feedback. Only then do our students benefit from the expertise of both their classroom teachers and the evaluators who support and coach them.

Problem Identification

The practice of using VAMs to evaluate student growth and educator effectiveness has become an inherent part of the culture of evaluation during the past fifteen years. A VAM provides a means to quantify educator effectiveness by assessing student academic growth against a prediction of expected growth based on previously collected data. This emerging evaluation model provides school leaders, parents, and students with precise information about how a student has performed in the past, how he/she is expected to perform on future assessments, and realistic achievement and growth goals. Figure 1.1 provides a sample VAM report from the Educator Value Added Assessment System (EVAAS), currently used in North Carolina.

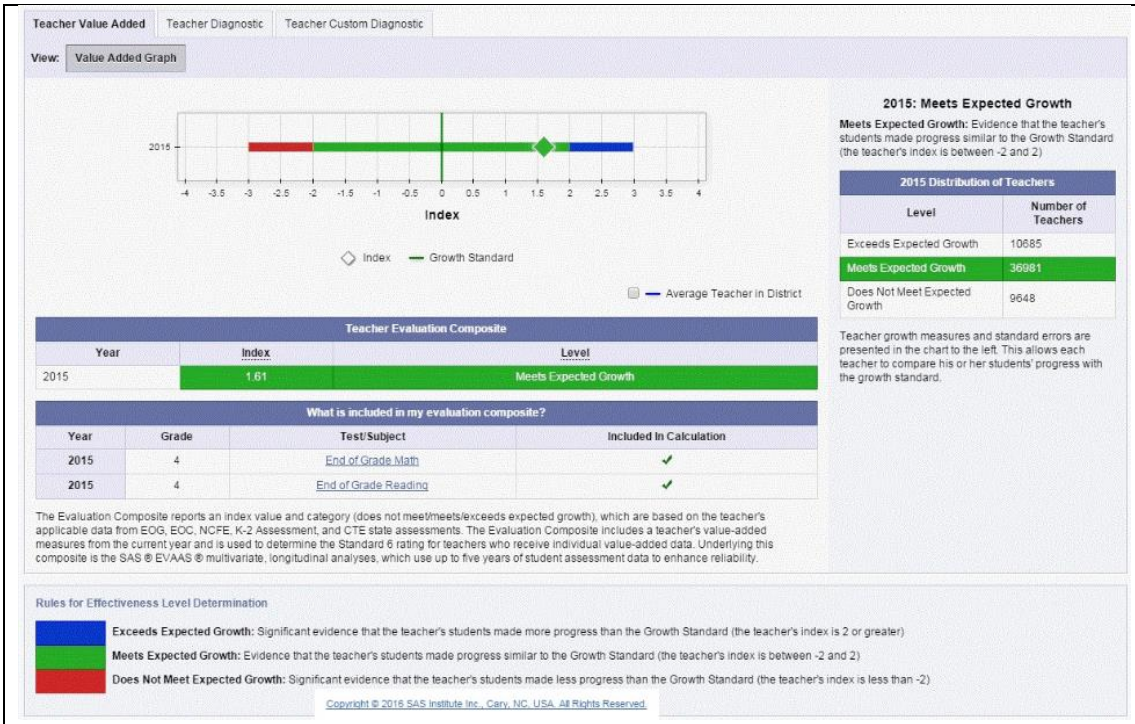


Figure 1.1. EVAAS Report Indicating a Teacher's Value-Added Measure Growth Score

Source: SAS provided permission for use of this image.
Figure 1.1 is divided into three specific sections. In the top of the report, a teacher's growth index is reported in a line graph. Red indicates that a teacher did not meet expected growth. Green indicates that a teacher did meet expected growth, and blue indicates that a teacher exceeded expected growth. In the middle section of the report, the teacher's index score for the 2014 school year is provided. The teacher's index was 0.39. One year of growth is indicated in EVAAS by an index of 0.00. Growth is determined within two standard errors of the mean (SEM). The third section of the report includes the teacher's courses included in the evaluation composite. The teacher can click on the links on each course to review growth data in each particular subject area. At the right of the report, teachers can see the distribution of teachers throughout the state.

The Impact of Value-Added Models on Evaluation

Unlike the proficiency model from the 2001 No Child Left Behind legislation where the goal was to ensure that all students were proficient on standardized assessments, the value-added era has focused more on realistic and achievable growth for

all students. However, VAMs have impacted teacher evaluation in ways that were not anticipated. Prior to the introduction of value-added models, No Child Left Behind (NCLB) legislation defined teacher quality through a designation of "highly qualified." Teachers could earn this distinction by having, at minimum, a bachelor's degree, a state license, and demonstrated competency in the course they taught (U.S. Department of Education, 2004). When researchers realized that these qualifications were not predictive of student learning outcomes on standardized assessments, the need to adopt a more conclusive method for determining educator effectiveness became a priority for states (Goldhaber, 2008).

The lack of correlation between how principals rate teachers and student learning outcomes led to a call for a new model of evaluating both student growth and educator effectiveness. North Carolina began using EVAAS in the mid-2000s to measure student growth and to provide predictions and projections for individual students based on historical standardized test data. However, in 2012, the state began using value-added data, as reported through EVAAS, to also evaluate educator effectiveness. As of the 2014-2015 school year, North Carolina used both ratings principal assigned and value-added scores to determine teacher effectiveness. Despite this effort to improve the assessment of teacher quality in North Carolina, an analysis of the state's evaluation data in both 2012 and 2013 revealed that North Carolina teacher evaluation ratings principals assign teachers show little correlation to value-added ratings teachers receive based on their students' growth (Tomberlin, 2014). Tables 1.1 and 1.2 provide the data that reveal the strong correlation between how evaluators rate teachers on Standards 1-5 of the Rubric for Evaluating North Carolina Teachers. The tables also show the low correlation

between those ratings and value-added ratings derived from how much growth students experience during the course of the school year (Standard 6).

Table 1.1

2011-2012 Correlations between Assigned Teacher Ratings and Value-Added Growth Ratings in North Carolina (Tomberlin, 2014)

	St. 1	St. 2	St. 3	St. 4	St. 5	St. 6
Standard 1	1.00	.715	.700	.721	.721	.186
Standard 2	.715	1.00	.711	.722	.696	.181
Standard 3	.700	.711	1.00	.758	.711	.204
Standard 4	.721	.722	.722	1.00	.714	.205
Standard 5	.721	.696	.711	.714	1.00	.173
Standard 6	.186	.181	.204	.205	.173	1.00

Table 1.2

2012-2013 Correlations between Assigned Teacher Ratings and Value-Added Growth Ratings in North Carolina (Tomberlin, 2014)

	St. 1	St. 2	St. 3	St. 4	St. 5	St. 6
Standard 1	1.00	.686	.665	.696	.695	.181
Standard 2	.686	1.00	.674	.715	.676	.167
Standard 3	.665	.674	1.00	.741	.682	.179
Standard 4	.696	.715	.741	1.00	.689	.198
Standard 5	.695	.676	.682	.689	1.00	.169
Standard 6	.181	.167	.179	.198	.169	1.00

In North Carolina during the 2011-2012 school year, the average correlation between and among ratings teachers received from evaluators on standards 1-5 was statistically significant at 0.71. However, the average correlation between how evaluators rated teachers and their value-added rating on standard 6 was not statistically significant at 0.18. The results from the 2012-2013 teacher evaluation data were similar. There was statistically significant correlation in how evaluators rated teachers on standards 1-5

(0.69) and again no statistical significance between how evaluators rated teachers and their value-added rating (0.17).

If evaluation ratings that principals assign to teachers are not aligned with student growth data, educational decision-makers must reconsider the process of teacher evaluation, the assignment of ratings, and the feedback teachers receive from evaluators. If evaluators rate teachers positively but the teachers' value-added data reveal insufficient student growth, teachers may not receive the constructive feedback that will lead to improvement in instructional practice. Furthermore, these same teachers may not sense the urgency to make changes to their practice based on their high marks on the five standards principals evaluate on teachers' summative assessments. By delving into the evaluation process, practitioners and researchers can gain a clearer understanding of how educators perceive the evaluation process, how well they understand the standards, whether or not they can identify effective teaching practices as specified by the indicators of each standard, and whether or not they know how to provide high-quality feedback to teachers.

Scrutiny of Value-Added Models

Currently, value-added models are the target of a great deal of national controversy, especially now that many states have recently adopted new, rigorous standards for students accompanied by new, challenging assessments. The increased rigor of both standards and assessments has resulted in "significant drops in the number of students reaching 'proficient' levels on assessments aligned to the new standards" (Value-Added Measures in Teacher Evaluation, 2014, paragraph 4). Furthermore, the American Statistical Association released a statement on using value-added models for

educational assessment in April 2014 that urges states against the use of VAM in evaluation (ASA, 2014). North Carolina contracted with West-Ed, an established consulting group with a mission to improve education through research, development and service, to compare VAMs for adoption in North Carolina. Based on their recommendations in a report titled, "Options for Incorporating Student Academic Growth Measure as One Measure of the Effectiveness of Teachers in Tested Grades and Subjects" from February 1, 2012, EVAAS emerged as a top contender in VAMs due to the use of both univariate and multi-variate statistical models, flexibility, features, and statistical performance (West-Ed, 2012). North Carolina selected EVAAS to serve as the vendor for the addition of VAMs to educator evaluation in the state. Value-added data is utilized to determine one-sixth of a teacher's evaluation in North Carolina currently. At this time, no legislation or state board policies have been written that inform dismissal, punitive action, or assignment based on a teacher's value-added effectiveness rating, although the original intention of using this model was to use the data to make employment decisions.

In November 2009, the Statistical Analysis System Institute (SAS) published a white paper entitled, *A Response to Criticisms of SAS EVAAS*. The paper addresses the different types of value-added models used by EVAAS – the multivariate (MRM) model and the univariate (URM) model. The paper addresses the following criticisms:

- Value-added models rely on standardized tests, which have limitations themselves.
- Missing student test data jeopardize the validity of analyses.

- Potential for rewards and punishments is related to class size (shrinkage estimation).
- SAS EVAAS does not adjust for socioeconomic factors.
- SAS EVAAS modeling lacks transparency and is too complex.
- SAS EVAAS statistical methods and algorithms have not been peer reviewed.
- SAS EVAAS predictions of student performance are not verified later.

Each criticism is met with a compelling argument including evidence of reliability, correlation, and sufficient stretch in the reporting scale of each assessment prior to accepting the assessment into the EVAAS value-added system (Sanders, Wright, Rivers, & Leandro, 2009). The authors contend that that EVAAS model has been peer reviewed, and they supply research from sources such as the U.S. Department of Education, National Center for Education Statistics, National Center on Performance Incentives at Vanderbilt University's Peabody College, Journal of Educational and Behavioral Statistics, and the American Educational Research Association to support the use of EVAAS. Regardless of the criticism around value-added measures, North Carolina has adopted the use of EVAAS as one of the components to determine educator effectiveness, and the model, notwithstanding of the scrutiny surrounding it, cannot be ignored in conversations about educator effectiveness in this state.

Introduction of a New Process for Teacher Evaluation in North Carolina

In August 2008, the North Carolina State Board of Education adopted a new evaluation rubric and process. Unlike its predecessor the Teacher Performance Appraisal Instrument – Revised (TPAI-R), the new Rubric for Evaluating North Carolina Teachers is far more robust with six standards, 25 elements, and 147 descriptors. Principals must

rate teachers on each of the five standards of the rubric based on thorough analysis of each of the elements and descriptors within that standard. Of the 25 elements, 17 are observable during classroom observations. During a classroom observation of a beginning teacher in his or her first three years of teaching, and all teachers on their renewal year (once every five years), a principal must complete three full formal observations where he/she rates the teacher on all 17 observable elements. During that observation, a principal must review 133 possible descriptors in the Rubric for Evaluating North Carolina Teachers. Each element is summarized in a paragraph. For each of the 25 elements in the rubric, a teacher must be rated as:

- Developing
- Proficient
- Accomplished
- Distinguished
- Not Demonstrated

Each descriptor is preceded by a detailed description of what an evaluator may see in a classroom that would warrant this specific rating. The tool is so comprehensive that the *North Carolina Teacher Evaluation Process Manual* is 50 pages long, and the rubric alone is 11 pages (Appendix A). Table 1.3 provides a breakdown of the number of elements and descriptors aligned with each standard in the Rubric for Evaluating North Carolina Teachers.

Table 1.3

Elements and Descriptors in the Rubric for Evaluating North Carolina Teachers

Standard	Number of Elements	Number of Descriptors per Element	Total Number of Descriptors Per Standard
Standard 1: Teachers demonstrate leadership	5	1a – 11* 1b – 8 1c – 6 1d – 4 1e – 4	33
Standard 2: Teachers establish a respectful environment for a diverse population of students	5	2a – 4* 2b – 8* 2c – 4* 2d – 8* 2e – 4	28
Standard 3: Teachers know the content they teach	4	3a – 8* 3b – 4* 3c – 8* 3d – 4*	24
Standard 4: Teachers facilitate learning for their students	8	4a – 7* 4b – 4* 4c – 4* 4d – 4* 4e – 11* 4f – 4* 4g – 8* 4h – 8*	50
Standard 5: Teachers reflect on their practice	3	5a – 4 5b – 4 5c – 4	12
Standard 6: Teachers contribute to the academic success of their students	N/A**		
Totals	25		147 133*

*Observable elements during classroom observations.
** Standard 6 is automatically populated by SAS EVAAS based on a teacher's value-added results from state standardized testing.

The Rubric for Evaluating North Carolina Teachers provides very clear, specific information regarding expected teacher performance. Because of the length of the rubric, the number of descriptors, and the subjectivity each evaluator brings into the principalship, the North Carolina Department of Public Instruction has provided state-wide and regional support with the process of evaluating teachers. Since the North

Carolina State Board of Education adopted the new Rubric for Evaluating North Carolina Teachers in 2008, the Educator Effectiveness division of NCDPI has provided state and regional support for interpreting and using the rubric. Nonetheless, regardless of training, because of the scope of the Rubric for Evaluating North Carolina Teachers and the lack of correlation between ratings principals assign to teachers and the value-added growth scores across the state, a more concerted effort must be made to address the quality of teacher evaluation across our state.

Educator Evaluation in Concordia Public Schools

Concordia Public Schools (CPS) is a small, urban-characteristic school district in the Northwestern foothills of North Carolina that serves approximately 3,200 students from two sister cities with a population of 21,000. Concordia is comprised of seven schools, including one comprehensive high school, one small problem-based learning magnet high school, one middle school, three elementary schools, and one public separate school for students with severe and profound disabilities. The district Concordia is 50% White, 24% Hispanic, 13% Black, 6% Multi-Racial, 6% Asian, < 1% American Indian, and < 1% Pacific Islander. Sixty-four percent of Concordia's students receive free or reduced lunch. Because of the district's size, central office staff and site-based administrators have opportunities to work together closely.

For school districts, the most logical place to begin a review of evaluation practices is at the local level. As Chief Academic Officer (CAO) for Concordia Public Schools, I began this review during my first year in the district. The educator evaluation data in CPS reflects the state trend. (The school district, schools, and educators have all been assigned pseudonyms.) Concordia Public Schools district proficiency data, EVAAS

data, comparison data between the district and state performance, and North Carolina Educator Evaluation System (NCEES) teacher evaluation data provide evidence that teacher evaluation is an area for improvement.

Framing the Problem

Complex problems often emerge not from one root cause but from a variety of causes that can be identified only through an examination of the problem through myriad lenses. Bolman and Deal (2008) contend that learning and applying multiple frames to an organizational problem is the best way to ensure the organization addresses all facets of the issue throughout the organization. This complex problem of practice in Concordia Public Schools can be framed through the cultural, structural, micro-political, and human resource lenses. This approach enabled my team and me to understand the complexity of the problem as well as to develop a plan for improvement.

Cultural Lens

The goals of central office-school partnerships include aligning the culture of the school to the district's mission, elevating the overall academic performance of each school, and providing ongoing, consistent support to each principal and school. According to Bolman and Deal (2008), "Culture forms the superglue that bonds an organization, unites people and helps an enterprise accomplish desired ends" (p. 253). Concordia Public Schools is a small district that spans only nine miles across one section of Concord County. The school district prides itself on its diversity and having a family-type atmosphere. However, the culture in each school varies from site to site and directly reflects each principal's leadership style, strengths, and emphases.

Cultural expectations were defined at the district level through collaborative development of mission, vision, and goals. The superintendent, district leaders, principals, assistant principals (APs), and instructional coaches worked in teams to help develop and provide feedback on the district's strategic plan, goals, and action steps during the 2013-14 school year. Goals and action steps were revisited during the summer of 2014, fall of 2014, and spring of 2015 in order to assess progress toward goals and determine next steps. District leaders also consistently emphasized data use in developing goals and worked closely with principals to develop next steps through weekly Executive Leadership Team (ELT) meetings, bi-monthly principals' meetings, and informal discussions.

Each district leader is assigned a school for each calendar school year, and that leader works closely with the principal and instructional staff of the school to support that school's improvement plan goals. These partnerships were developed by the superintendent based on areas of expertise, personality, and the needs of each school. The district leader completes weekly collaborative classroom walkthroughs with a site-based leader and meets with the principal regularly to provide support and guidance.

The organizational culture of the district is evident in structures that exist to promote transparency. Weekly three-hour ELT meetings include the superintendent, assistant superintendent, chief academic officer, director of finance, director of human resources, director of elementary education, director of high schools, and director of accountability and technology. Principals, assistant principals, and instructional coaches are on a rotating schedule to attend ELT. All shared notes from the meeting are accessible online for all site-based leaders, and the site-based representative of the week

takes shared notes for all principals, APs, and coaches to review. Furthermore, a representative from the group captures the key information from the meeting in a Google document that is e-mailed to all faculty and staff members in the district at the end of each weekly ELT meeting. The goal of this sharing of information to all stakeholders is to promote a culture of transparency and trust.

Bi-monthly principals meetings provide another opportunity for site-based leaders and district-level leaders to collaborate and communicate in both large and small groups. Features such as the "un-meeting," a 30-minute standing agenda item where principals have the opportunity to share concerns or ask for support from other site-based or district-level leaders, and "work with director time" are built into principals' meetings give principals an opportunity for support from both the collective group and from their central office partner. This time is designed to build trust as well as capacity. By providing opportunities for candid discussion, district leaders hope that concerns around issues such as evaluation will emerge.

One facet of developing an organizational culture is promoting an atmosphere of honest dialogue where site-based leaders feel comfortable discussing classroom practice and learning from one another. During the 2013-14 school year, district leadership instituted a classroom walkthrough tool (CWT) aligned to district instructional goals (Appendix B). Administrators complete partner walkthroughs weekly and small-group collaborative walkthroughs monthly. The goal of developing an expectation around weekly and monthly collaborative walkthroughs is not only to gauge instruction in classrooms but also to build a sense of camaraderie and mutual support in order to build an organizational culture of collaboration and trust among principals and district leaders.

Weekly walkthroughs also keep directors and site-based administrators in tune with instruction taking place in classrooms. Walkthroughs support the collaborative culture in CPS by providing leaders the opportunity to build a common language and expectation of instruction.

Structural Lens

Both the structure of Concordia Public Schools' administrative staff and the expectations of the principalship in general have shifted in the past few years. In the past three years, principals in Concordia Public Schools have been charged with learning new content standards, changing online platforms for the North Carolina online submission of ratings in the Rubric for Evaluating North Carolina Teachers twice, and helping teachers understand the implications of a new value-added evaluation model. These changes have put a great deal of new learning on the shoulders of principals. Two principals were hired in July 2013, two changed schools, and three moved their staffs into new locations. Moreover, principals in CPS have had to adjust to a new central office administrative team, complete with new superintendent, assistant superintendent, chief academic officer, and three directors. Principals in Concordia Public Schools have been faced with structural changes to the way they evaluate, the standards they evaluate, the platforms they use to evaluate, and a new expectation from district leadership.

In order to meet the needs of principals, the district considered high-quality, ongoing professional development to help them acquire the requisite skills needed to evaluate accurately and provide teachers with necessary feedback. Guskey (2009) relates that "effective professional learning time must be well organized, carefully structured, clearly focused, and purposefully directed" (p. 230). Unfortunately, in an age of high-

stakes accountability, new standards, and fewer financial resources, district leaders are doing more with less, including less personnel, to plan and implement professional development. One of the major concerns in CPS is that cutting instructional positions, such as the instructional coach at the middle school and assistant principals at the elementary schools, has doubled the work of the principal.

High quality professional development can be characterized by including a strong focus on content, providing opportunities for active learning, developing a sense of coherence, having adequate duration to accomplish the goals of the training, and providing opportunities for collective participation (Desimone, 2009; Yoon, Duncan, Lee, Scarloss, & Shapley, 2007). Instead of automatically adopting topics for training based on current trends and popularity, district leaders must analyze relevant staff, student, school, and district data in order to design appropriate and quality professional development (Roy & Hord, 2003). In planning for this professional development for evaluators in CPS, a great deal of emphasis was placed not only on the needs of the principals from the viewpoint of district leaders but also the training principals themselves identified as a need. The superintendent and district leaders took the approach with principals that "we're all learning this together" to build trust as well as capacity in the district.

Micro-Political Lens

The lack of correlation between teacher evaluation ratings and value-added scores prompted my colleagues and me to conduct an analysis of current state policies. This analysis revealed that North Carolina has emphasized procedures and tools in evaluation policies but has not set forth any precedent through policy about how evaluators should

be trained on the content of the evaluation instrument or the crucial coaching sessions that occur after each observation and at the end of the year during a teacher's summative evaluation conference. As classroom culture changed at the onset of the twenty-first century, the State Board of Education, under the advisement of the North Carolina Department of Public Instruction (NCDPI), developed a commission to create new North Carolina Professional Teaching Standards, which resulted in a new, more complex evaluation instrument – Rubric for Evaluating North Carolina Teachers. The standards and their descriptors were vastly divergent from the previous evaluation instrument used in the state – the TPAI-R. The novelty and complexity of the new tool resulted in a decision to spend four years (from 2007-2011) to complete the full implementation of the new standards. This emphasis on the new evaluation process as well as North Carolina's first online reporting system, through Mid-Continent Research for Education and Learning (McREL), proved to be challenging for many evaluators, and a concerted effort went into ensuring that training was provided on the evaluation process through the Educator Evaluation division at NCDPI.

Educator Evaluation consultants provided face-to-face and online professional development as well as a comprehensive wikispace where resources, archived webinars, and other support materials are housed. The NCEES Wikispace is devoted entirely to evaluation in North Carolina (Appendix C). However, the majority of the trainings and webinars were developed to support evaluators with necessary operational elements of the system such as logging into the online system, navigating the platform, and use of the online system. These aspects of training were necessary to comply with the evaluation

process and they overshadowed the equally-significant shift of ensuring target agreement and a deep understanding of the standards.

North Carolina has only three state board policies regarding evaluation. *TCP-C-004 –Policy Establishing the Teacher Performance Appraisal Process* (Appendix D) ensures principals and teachers are trained on the process and guidelines of evaluation each year. This policy ensures that principals and teachers understand the timeline for teacher completion of the Professional Development Plan (PDP) and when evaluations will take place. The policy also provides direction so that certifying teachers understand whether they are on a full-observation cycle or an abbreviated cycle. A teacher's cycle is determined by his or her license renewal. Every five years, teachers must renew their North Carolina teaching licenses. In terms of evaluation, teachers in their fifth year are observed three times during the school year for a minimum of 45 minutes or a full class period, whichever is longer. These three observations are provided as formative assessment for teachers to give them an indication of what the evaluator observes. Evaluators conference with teachers and provide opportunities to grow and improve throughout the year. However, teachers who are not in their year of license renewal are evaluated on the abbreviated renewal cycle. These teachers receive only three 20-minute "snapshot" observations over the course of the year as formative assessment on only Standards 1 and 4. Furthermore, *TCP-C-004* also provides guidelines for pre-observation and post-observation conferences. A pre-observation conference must take place prior to the first observation of the year for teachers in their renewal year, and a post-observation conference must occur within ten days of any observation, full or abbreviated, or the observation is void. These stipulations are clearly defined by this policy.

TCP-C-006 – Policy on Standards and Criteria for Evaluation of Professional School Employees – establishes the six standards on which educators are evaluated (Appendix E). The standards fall into two categories – those which are evaluated by the principal or her designee and one that is populated based on value-added growth data from student summative assessment data. The North Carolina Professional Teaching Standards include: Standard 1: Teachers Demonstrate Leadership, Standard 2: Teachers Establish a Respectful Environment for the Diverse Needs of Their Students, Standard 3: Teachers Know the Content they Teach, Standard 4: Teachers Facilitate Learning for Their Students, Standard 5: Teachers Reflect on Their Practice, and Standard 6: Teachers Contribute to the Academic Success of Their Students. Standards 1-5 are evaluated by the principal, and Standard VI is populated from value-added data. A more comprehensive breakdown of the standards and each of the 25 elements can be found in the *North Carolina Teacher Evaluation Process Manual*. The Rubric for Evaluating North Carolina Teachers serves as a growth model. By breaking down the responsibilities of teachers into 25 clear and specific elements, administrators can work directly with teachers to pinpoint strengths and areas for improvement.

TCS-C-021 – Policy on Educator Value-Added Assessment System (EVAAS) Teacher Module likewise provides information about the Board's decision to adopt a value-added model to evaluate educator effectiveness (Appendix F). North Carolina currently has no policies regarding evaluation training or certification for any evaluator. Furthermore, principal preparation programs do not have a common, shared requirement for evaluator training. It may be argued that capacity building is too costly and is not needed, since all principals are required to study teacher evaluation during their licensure

preparation. However, how principals develop their evaluation skills and gain insight into how to provide appropriate feedback is dependent upon the institution of higher education (IHE) and not defined by the state. This lack of consistency in principal preparation programs may contribute to the problem currently experienced in Concordia Public Schools.

Human Resource Lens from the State Level

In North Carolina's public school districts, evaluation is located under the auspices of human resources. Human resources is responsible for enforcing policies related to evaluation, supporting principals with the documentation and process for working with ineffective teachers, and for serving as a liaison between the schools and the North Carolina Department of Public Instruction. Although human resource leaders enforce policies and provide evaluation support to principals, there may be a need to provide more focused support through professional development and capacity building from local school districts.

Ideally, teachers use the ratings and feedback from principals to improve their practice. However, if ratings are not aligned with value-added measures, teachers may receive conflicting information about the quality of their practice and may not receive the appropriate feedback to improve. Current state support focuses mostly on policies regarding evaluation set forth by the state rather than around understanding the standards, target agreement, and coaching conversations with teachers. The lack of correlation reported by Tomberlin (2014) between the ratings principals assign teachers and teachers' value-added scores may indicate that practicing principals lack the requisite knowledge and skills to evaluate teachers and need further training.

Human Resource Lens from the Local Level

During the 2013-14 school year, district-level leadership at CPS provided minimal support to evaluators regarding the educator evaluation system beyond the mandatory timeline review as specified in *TCP-C-004*. Evidence of this lack of support in even the basic elements of evaluation was revealed by the fact that one first-year principal was not aware that some teachers were on the abbreviated cycle of evaluation, while others were on the renewal cycle. This particular administrator completed the full-scale renewal cycle evaluation process on each teacher in her building, evaluating all five standards and conducting three full-length classroom observations when the majority of the faculty was actually on the abbreviated cycle and could have been evaluated on standards 1 and 4 only, with three abbreviated evaluations.

New site-based administrators need a great deal of support from the district's human resources leaders in order to understand policies relating to the evaluation tool but also to gain a deep understanding of how to use the tool to support, monitor, direct, and appropriately evaluate staff. Veteran principals also need refresher training to review the standards and engage in discourse regarding the "look-fors" for each of the 17 observable elements. District leadership is responsible for creating opportunities for all principals to engage with the standards in multiple ways.

During the 2013-14 school year, two principals began their first principalships, two principals were transferred to new schools, and three experienced principals remained at the schools they had led the previous year. Also during this year, the district hired a new superintendent, assistant superintendent, chief academic officer, director of elementary education, director of exceptional children, and director of accountability and

technology. With such a significant turnover in leadership, the 2013-14 school year was one of adjusting to new roles and a new leadership structure as well as trust building. However, it was also a year of observing and evaluating to determine areas of strength and areas for growth. New central office leaders visited schools to learn more about their culture and achievement, as well as to build relationships and coach site-based administrators. These observations revealed discrepancies between observed instruction and ratings assigned to teachers on the Rubric for Evaluating North Carolina Teachers.

In terms of the human resource lens, a significant consideration is the lack of sufficient personnel. Staffing and support from central office leadership are important considerations. Over the past several years, funding for North Carolina Public Schools has declined significantly, around $200 million since 2008. Many administrative and support positions have thus been cut to preserve classroom teaching positions (Public School Forum of North Carolina, 2013). In CPS, elementary school assistant principals, one middle school instructional coach, and a district-wide instructional technology facilitator have all been cut to preserve classroom teaching positions. Bolman and Deal (2008) point out that, "Emerging evidence suggests that downsizing has often produced disappointing results" (p. 138). One aspect to consider is how the loss of these key, instructionally-focused support positions has impacted teacher evaluation.

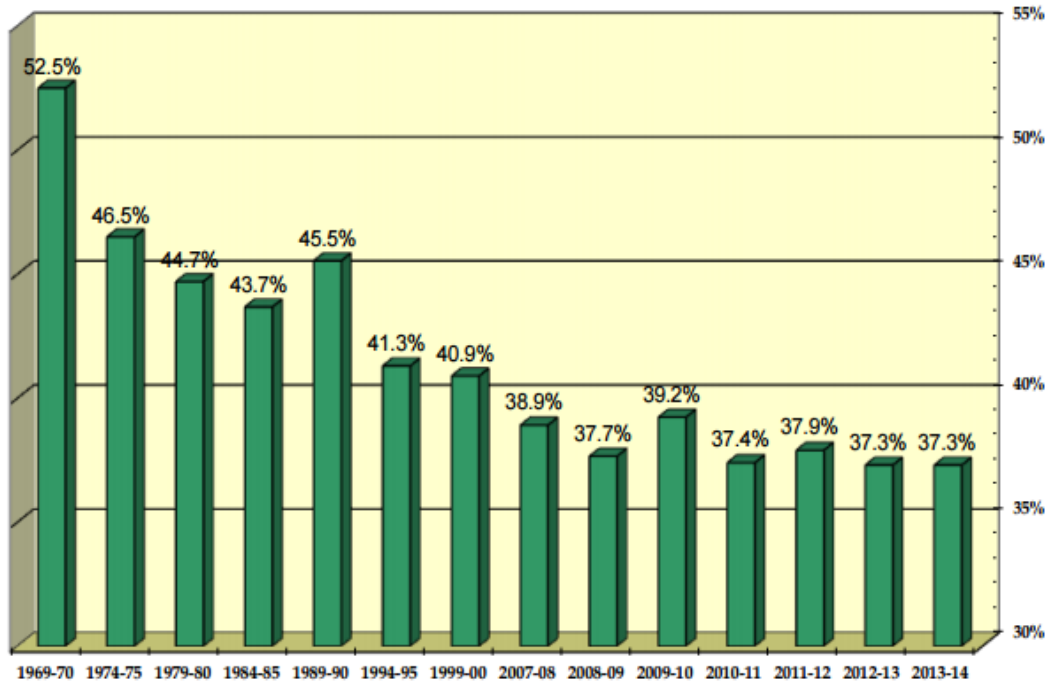


Figure 1.2. Percent of the General Fund Public Schools Appropriations

Source: Highlights of the North Carolina Public School budget (p. 2, Rep.). (2014). Raleigh, NC: North Carolina Department of Public Instruction.

Another example of the lack of robust support from the district level is that two different principals asked for assistance when placing teachers on "monitored improvement plans," previously referred to as "action plans." Both principals were provided with a locally-developed action plan template from the human resources department and were not provided with the procedures for formal implementation of these plans through NCEES. At the end of the year, the plans they created for these teachers could not be recognized by the state because the proper procedures, digital forms, and documentation were not entered into the NCEES.

In an attempt to learn more about how principals evaluate at the local level, I interviewed an experienced assistant superintendent of human resources in Appalachian County Schools (ACS) (pseudonym) (personal communication, June 21, 2014). ACS was one of two districts to participate in the NCDPI Observation Calibration Training (OCT) Mini-Pilot in July 2014. This pilot was designed to improve evaluator accuracy and reduce bias by providing participants with the opportunity to rate teachers in videos housed online and receive immediate feedback. During the same time period, the NCDPI NCEES Consultant who oversaw the OCT Mini-Pilot, interviewed the director of human resources in another LEA using the same interview questions (personal communication, June 2014). Interviews were coded using the Provisional Coding model. A list of codes was developed, modified, and revised based on the content of the interviews.

Several themes emerged in these interviews regarding the realities of principal preparation to evaluate teachers in North Carolina school districts. The interviews corroborated the team's assumptions regarding the lack of correlation between principal ratings and value-added ratings. Both human resource experts cite the following as problems with the current state of evaluation in North Carolina:

- General lack of content and pedagogical knowledge
- Lack of understanding of the evaluation tool/standards
- More support needed from local education agency (LEA)
- Inability/lack of motivation to deal with conflict
- Budget concerns in the state led to fear of losing positions
- Changing online platforms twice in three years has forced a time commitment on procedure, not evaluation itself

- Incorrect/improper training

The NCEES Consultant indicated that these findings mirrored those she had uncovered from across the state in data she had collected from quarterly Principal READY meetings held regionally for principals across North Carolina (personal communication, July 7, 2014). The interviews, data provided by Simmons, and other informal conversations with members of the improvement team, principals, the superintendent, and the School Board attorney also illustrated more succinct insight into the LEA-level issues in Concordia and other districts.

District-level support for principals, in terms of human resources, has generally been provided from a task-oriented standpoint – that is, reminding principals of due dates, providing changes in policies, and working with personnel issues. However, process issues are seldom discussed in terms of evaluation, and evaluators need opportunities to focus on the evaluation tool, pedagogy, content, coaching teachers, and dealing with conflict. Furthermore, not all directors of human resources have a deep understanding of evaluation, the NCEES, or coaching teachers. Gandha and Baxter (2015) agree. "The 'train the trainer' model is popular but could inadvertently contribute to implementation inconsistency" (p. 8). The Southern Regional Education Board endorses the use of standardized training materials, taped trainings and webinars, supplemental web-based resources, and ongoing communication as "key strategies employed by SREB states to improve the consistency and quality of local training while managing cost" (Gandha & Baxter, 2015, p. 8).

Challenges and Problems in Evaluation Practices

An analysis of the literature related to teacher evaluation indicates the impact of inaccurate evaluation practices have on schools and districts. With both proficiency and EVAAS data at educators' disposal, we have more data than ever before to pinpoint areas of specific teachers' strengths and areas for improvement. However, these data are of little use without leaders who can accurately evaluate the data to provide high-quality feedback. Instructional guidance, support, and feedback that teachers receive from principals is imperative in improving their practice. Research has indicated that evaluation is more effective when the evaluators are trained (Darling-Hammond, Amrein-Beardsley, and Rothstein, 2011). Training should include resources that support the evaluation process (McGuinn, 2012). Accurate, high-quality evaluation should lead to feedback on instructional practices which can provide teachers with necessary information to improve the quality of instruction. This link between evaluation and feedback to improve teaching is significant because for almost two decades, quality teaching has been consistently identified by researchers as the most important school-based factor in student achievement (McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Rivkin, Hanushek, & Kain, 2000; Rowan, Correnti & Miller, 2002; Wright, Horn, & Sanders, 1997). Because improving teacher evaluation can lead to a ripple effect that can improve student learning outcomes by way of constructive feedback and changes in instructional practice, the worthy goal of this project is to improve evaluation practices in order to impact student learning outcomes.

The Widget Effect in Evaluation

The "widget effect," reported by Weisburg et al. (2009), describes a school district's assumption that teacher effectiveness is the same from teacher to teacher. The report suggests that evaluators rate all teachers the same. The widget effect suggests that teachers are viewed as interchangeable parts, not as individuals. This report suggests that better evaluation will not only improve teaching to benefit students but will also benefit teachers by treating them as professionals. Key findings of the widget effect in teacher evaluation include:

- All teachers are rated good or great;
- Excellence goes unrecognized;
- Inadequate professional development is provided;
- No special attention is given to novice teachers;
- Poor performance goes unaddressed.

These findings serve as another indicator that principal instructional leadership has taken a back seat to managerial and organizational components of the principal's role. Without a clear emphasis from the state on evaluation, ongoing opportunities for discourse, high-quality professional development, and local and state support, the quality and accuracy of teacher evaluations is not likely to improve.

As more and more states are turning to measuring student growth data as using VAM such as the EVAAS, the widget effect is more prominent than ever. VAM data have revealed notable discrepancies between evaluation and student learning outcomes. Value-added assessment systems, such as EVAAS, provide individual teacher, school, district and state growth data. EVAAS also determines the effectiveness of teachers,

schools, and districts regarding student achievement and provides multiple reports aimed at analyzing student and teacher performance on standardized assessments.

The Widget Effect in North Carolina

Research by Batton, Britt, DeNeal, & Hales (2012) provides data that support the presence of the widget effect in North Carolina. A multifactorial correlational study was conducted to analyze the data surrounding the correlation between teacher performance evaluation ratings and EVAAS student achievement date. The dataset consisted of 11,430 North Carolina teachers in 35 local education agencies (LEAs) having both EVAAS scores and performance evaluation ratings assigned in 2010-11 school year. Although 46,000 teachers had evaluation data for 2010-11, only those 11,000+ included an end-of-grade (EOG) or end-of-course (EOC) assessment. This group received an EVAAS data score. Researchers found that there was only a small distribution of evaluation ratings in the study. Of the group, the 100 teachers who contributed to the greatest student growth received virtually the same ratings (3.8) as the 100 teachers who contributed to the least student growth (3.2). According to the study, both sets of teachers were rated *Proficient* or *Accomplished* by their principals on their performance evaluation ratings. The study did not find a correlation between performance evaluation data and EVAAS data (Batton et al., 2012). This finding alone demonstrates a need for better preparation for our state's teacher evaluators. In order to meet this need, a comprehensive system of support from district and state agencies is mandatory.

In their SREB Report, Gandha and Baxter (2015) relate that researchers have discovered that observers' judgments are often compromised by their experiences, beliefs, prior knowledge, and biases. They contend that observers' expectations could

"change with time (drift) as observers consciously or subconsciously adjust their expectations" (p. 10). This recognition of the challenges observers and trainers face provides even more impetus for a change in the way North Carolina supports and trains both evaluators and those who train the trainers. This research, coupled with clear data from North Carolina's EVAAS results, suggests that evaluators need training on accuracy and calibration.

Instructional Leadership and Evaluation

The "principal as instructional leader" is a concept that emerged during the early 1980's. Prior to that, most principals functioned as managerial and operational leaders. Research during that time indicated that effective schools most often had principals who understood and articulated the importance of instructional leadership (Brookover & Lezotte, 1982). Today, instructional leadership is still a critical component of the principalship. Evaluators must be able to measure instruction accurately during classroom observations. "Classroom observation is a powerful component of teacher evaluation systems. It measures instructional practice, provides clarification on what effective teaching looks like and gives teachers the concrete and actionable feedback they need to improve teaching practice" (Gandha & Baxter, 2015, p. 3). Skilled evaluators are essential to a valid and reliable educator evaluation process. Until the emergence of value-added models, which provided data to support or refute teacher effectiveness, objectivity in evaluation was even more difficult to measure. This shift to accessible data that clearly and accurately reveal whether or not students demonstrate growth calls into question the subjective notion of evaluation. EVAAS provides, without bias, data regarding whether or not students grow in their acquisition of content knowledge and

understanding based on their performance on the North Carolina EOGs and EOCs. With widespread access to this data, public education suddenly became much more "public."

One way principals exhibit instructional leadership is through the process of teacher evaluation. The multi-factorial correlation study findings by Batton et al. (2012) indicate that principals rate the most effective teachers virtually the same as they do ineffective teachers. This phenomenon supports the notion that all teachers are rated as good or great. If teachers are rated the same, regardless of their effectiveness with students, then teachers are not getting the feedback they need to improve instruction. Both state and district leaders must work together to implement a sustainable plan to give teacher evaluators the requisite knowledge and skills to evaluate and provide feedback to ensure teacher growth and effectiveness. At the district level, requirements include an institutional emphasis on evaluator accountability, high-quality professional development, ongoing dialogue, and support.

However, state policy changes and support protocols alone will not ensure that evaluators rate teachers appropriately or that their feedback is clear, specific, and appropriate. "Whether they currently assess observers or not, states uniformly agree that assessment and certification is not a substitute for continuous observer training and calibration" (Gandha & Baxter, 2015, p. 10). Much of this responsibility falls to district-level leaders. Numerous studies provide evidence that a combination of rigorous classroom observations and additional data measures will provide an accurate evaluation of teacher effectiveness (Bill & Melinda Gates Foundation, 2013; Ho & Kane, 2013; Taylor & Tyler, 2012). Providing the type of training necessary to ensure accurate evaluation is often beyond the scope of what district-level support can deliver in a small

district. An explicit state-district partnership is needed to improve the quality of evaluation of North Carolina's teachers. State agencies must provide high-quality training, tools, and support that focus more on developing target agreement, scoring bias, and instructional coaching. The district level must provide additional support through evaluator accountability, high-quality professional development, and ongoing dialogue. This partnership is paramount to ensuring not only that teachers receive clear, specific feedback to improve their practice but also that evaluator ratings correlate to EVAAS value-added results.

This improvement project incorporates a state-district partnership to address observation calibration. Gandha and Baxter (2015) report that "Observers could also improve their rating quality through ongoing calibration opportunities" (p. 12). They contend that calibration might include watching pre-taped video segments that have been pre-rated by master raters. This would provide evaluators with the opportunity to observe the same teacher and engage in discussion regarding their notes, ratings, and the language of the standards. The goal is to provide evaluators with "multiple and ongoing opportunities to reflect on their rating accuracy and the basis on which they evaluate teaching, and to increase awareness of potential systematic biases influencing their judgments" (p. 12). This project was developed around the notion that the power of prolonged engagement is key in terms of improving evaluation practices.

CHAPTER 2: HISTORY AND REVIEW OF THE PROBLEM

For most policy makers and district leaders, one essential determinant of teacher quality is teacher evaluation (Simmons, & Mullins, 2013). High-stakes decisions, such as staff retention and performance pay are becoming inextricably bound to evaluation results. However, recent reviews of teacher evaluation data in North Carolina have revealed discrepancies between the ratings principals assign to teachers on their evaluations and the value-added ratings they receive based on the amount of positive or negative growth of students they teach. "The ongoing challenge for many states is developing an accurate understanding of different levels of teaching quality that is shared by all educators" (Gandha & Baxter, 2015, p. 7). According to Tomberlin's (2014) research, in North Carolina there is no correlation between ratings principal assign to teachers and value-added effectiveness ratings teachers receive based on student assessment data. Therefore, determining a teacher's true effectiveness has become elusive. Furthermore, the lack of congruence between teacher performance and principal evaluation of teachers has been targeted by the North Carolina Department of Public Instruction as a significant problem. An evaluation of the Concordia Public Schools district proficiency data, EVAAS data, comparison data between the district and state performance, and NCEES teacher evaluation data provide quality evidence that teacher evaluation is an area for improvement.

Findings from Data

In January 2015, I performed Spearman-R correlations to assess whether the ratings seven principals in Concordia Public Schools assigned to teachers on the North

Carolina Educator Evaluation System correlated with teacher value-added ratings derived from student growth data. Ratings principals assigned to teachers were obtained from the North Carolina Educator Evaluation System. Ratings of Developing (1), Proficient (2), Accomplished (3), and Distinguished (4) are assigned to teachers on five different standards:

- Standard 1 – Teachers demonstrate leadership;
- Standard 2 – Teachers establish a respectful learning environment for a diverse population of students;
- Standard 3 – Teachers know the content they teach;
- Standard 4 – Teachers facilitate learning for their students;
- Standard 5 – Teachers reflect on their practice.

Ratings for Standard 6, value-added growth ratings, were derived from data provided by SAS through EVAAS are: Does not meet expected growth (1), Meets expected growth (2), or Exceeds expected growth (3). Teacher ratings were collected from 119 teachers who also received an EVAAS value-added rating based on student growth. Statistically significant correlations exist between each of the five ratings principals assign teachers. Each of these correlations is significant at the 0.01 level (two-tailed). The strongest correlations exist between Standard 3 – Content Knowledge and Standard 4 – Pedagogy (.571), Standard 1 –Leadership and 4 -Pedagogy (.558), and Standard 2 – Climate and Culture and Standard 5 – Reflection (.513). Correlations between Standard 6 – Value-Added Effectiveness and each of the principal assigned standards 1-5 are far less statistically significant. The most significant correlation between Standard 6 and any other standard is Standard 3 – Content Knowledge (.227). The lowest correlation

between Standard 6 and another standard is Standard 5 – Reflection (.113). The six Spearman-R correlations are reported in Table 2.1.

Table 2.1

2013-2014 Correlations between Assigned Teacher Ratings and Value-Added Growth Ratings in Concordia Public Schools

	St. 1	St. 2	St. 3	St. 4	St. 5	St. 6
Standard 1	1.00	.461**	.432**	.558**	.406**	.165
Standard 2	.461**	1.00	.470**	.499**	.513**	.115
Standard 3	.432**	.470**	1.00	.571**	.431**	.227*
Standard 4	.558**	.499**	.571**	1.00	.396**	.199*
Standard 5	.406**	.513**	.413**	.396**	1.00	.113
Standard 6	.165	.115	.227*	.199*	.113	1.00

*Correlation significant at the 0.05 level (2-tailed)
**Correlation significant at the 0.01 level (2-tailed)

These findings from Concordia Public Schools mirror the findings of Tomberlin (2014) in a study of all North Carolina teachers who received a Standard 6 rating based on student value-added data during the 2012-2013 school year. In both studies, there is a statistically strong correlation among ratings principals assign teachers in standards 1-5. In other words, if a principal assigns a teacher a rating of "proficient" for one standard, there is a strong probability that the teacher will receive a rating of "proficient" on the remaining standards. However, there is little correlation between how a principal rates teachers on standards 1-5 and the rating a teacher receives from his/her value-added rating on standard 6.

In Concordia Public Schools, although 25% of teachers did not meet expected growth based on their 2013-2014 value-added EVAAS value-added growth rating, all teachers who received a rating of "did not meet expected growth" were rated by

principals as “proficient,” “accomplished,” or “distinguished” on all five standards of the Rubric for Evaluating North Carolina Teachers. During the 2012-2013 year, fifteen teachers in CPS did not meet expected growth. Each of those teachers was rated by their principals as “proficient,” “accomplished,” or “distinguished” on NCEES standards 1-5. During the 2013-2014 school year, thirty teachers in CPS did not meet expected growth; likewise each teacher received ratings of “proficient,” “accomplished,” or “distinguished” on all ratings on standards 1-5 assigned by principals. These findings indicate a district-wide trend of a consistent lack of correlation between teacher evaluation and student achievement and a trend in the district.

The teacher evaluation process in North Carolina is a growth model. One expectation of the model is that at least some teachers will be rated as “developing” on some of the five standards. During the 2013-14 school year, 189 teachers were evaluated who returned to Concordia Public Schools for the 2014-15 school year. Of those 189 teachers, no teacher was rated as “developing” on any of the five teacher evaluation standards assigned by principals. Furthermore, according to the director of human resources, only one of the teachers who left the school district after 2013-14 received a rating of “developing” on any standard on his/her summative evaluation. However, based on 2013-14 EVAAS data, 25% of teachers in Concordia Public Schools did not meet expected growth on standard 6, based on teacher’s value-added data released in October 2014. The state average for teachers who did not meet expected growth is 15%. This 10% difference in the number of teachers who did not meet expected growth is an area for concern. Based on the same preliminary EVAAS data, only 56.5% of Concordia Middle School (CMS) teachers met or exceeded expected growth on their standard 6

rating. Again, no teacher at CMS was rated lower than proficient on any standard 1-5. Furthermore, five of the teachers who did not meet expected growth based on EVAAS data received at least one rating of "accomplished," and one of the teachers received two ratings of "distinguished" and three ratings of "accomplished" on standards 1-5. These data provide two layers of evidence that this problem of practice is pervasive at Concordia Public Schools

Comparative Proficiency Data

Another indicator of the severity of the problem is found in comparative proficiency data. During the 2011-2012 and the 2012-2013 school years, North Carolina assessed student achievement in seven subjects in elementary schools, seven subjects in middle schools, and three subjects at the high school level. From 2012-2013 to 2013-2014, students in Concordia Public Schools experienced a significant decline in their performance in terms of proficiency when compared with the state average. In 2011-12, the district's elementary schools exceeded the state average in proficiency in 71.4% of tested subjects. During that same year, the middle school exceeded the state average in 100% of tested subjects. North Carolina adopted new content-area standards for all content areas during the 2012-13 school year, and proficiency dropped across the state. However, Concordia fell below the state average in 71.5% of tested subjects in elementary school and 100% of tested subjects in middle school. However, teacher evaluation ratings indicated that every teacher who taught these tested subjects was rated as "proficient," "accomplished," or "distinguished" on each of the five standards assigned by principals. These data indicate one discrepancy between performance on state assessments and teacher ratings.

These findings led to a deeper investigation of evaluation practices. These data led me to contact the NCDPI NCEES Consultant and began discussions regarding how the district could partner with the state to improve principal practice and to gain further insight into why principal ratings did not correlate with value-added ratings. Conversations with the state consultant led to Concordia's participation in the NCDPI OCT Mini-Pilot, which served as a first step toward implementing a formal intervention and assessment plan to address the lack of correlation between principal ratings and teacher value-added growth ratings in the NCEES. The OCT Mini-Pilot served as the NCDPI response to requests from evaluators across the state to have access to an online, self-paced observation training tool. The online tool gives North Carolina evaluators an opportunity to observe classrooms through video, rate the teachers in the videos using the Rubric for Evaluating North Carolina Teachers, and receive immediate feedback on the accuracy of their ratings.

Current and Desired States

Current State

Concordia Public Schools district proficiency data, EVAAS data, comparison data between the district and state performance, and NCEES teacher evaluation data provide multiple sources of evidence that teacher evaluation is an area for improvement. Principals' evaluation of teachers and value-added data reveal that CPS fell below the state average in 69.3% of tested subjects. In 2013-14, 25% of CPS teachers did not meet expected growth (as indicated by EVAAS data reported from state summative assessments), while statewide only 15% of teachers did not do so. In addition to poor student performance and less growth than expected across the district, principals rated

every teacher in the district a minimum of "proficient" on each standard of their evaluations.

Experience and compliance as factors for consideration. Low student performance and principal lack of compliance with evaluation protocols appear to be related to level of principal experience. CPS hired two inexperienced, first-year principals in new positions outside their experience levels. The two schools with new principals were the two lowest-performing schools in the district, according to North Carolina standardized test data. Mr. Black, who completed his first year at Northeast Elementary School during the 2013-2014 school year, has a background as a high school teacher and middle school assistant principal, while the new middle school principal, Mrs. Howard, has spent her entire 20+ year career in high schools. In addition, Mr. Black did not complete his evaluations in a timely manner, and central office staff as well as other licensed evaluators in the district came into the school in the late spring of 2014 to help complete the observations so that the principal would be in compliance with state law. This issue resulted in teachers not receiving timely feedback in order to make adjustments to their instruction. Interestingly, Northeast Elementary was the second lowest performing school in CPS during the 2013-14 school year.

Theoretical Framework

Currently, North Carolina's Department of Public Instruction offers a variety of support for principals including bi-annual Principal READY meetings for all North Carolina principals, Principals' Council meetings for a representative sample of principals in each region, synchronous and asynchronous online resources and webinars, support through the NCEES wiki, and access to the North Carolina Educator Evaluation

Consultant for virtual or face-to-face support. At the district level, human resources directors, curriculum leaders, and superintendents all have an impact on priorities for principals, their professional development, and their growth.

The theoretical framework guiding this project is grounded in a modification of the synthesis of research on the "ripple effect" by Clifford, Behrstock-Sherratt, & Fetters (2012) with emphasis on the professional development framework developed by Thomas Guskey (2000). According to Hargreaves and Fink (2006), "What leaders do in one school necessarily affects the fortunes of students and teachers in other schools around them; their actions reverberate throughout the system like ripples in a pond" (p. 16). This metaphor expands beyond the classroom. Our notion is that the ripple effect exists on a larger scale in the greater educational system. Marzano, Waters, and McNulty (2005) concluded that "a highly effective school leader can have a dramatic influence on the overall academic achievement of students" (p. 10). In terms of evaluation, both state structures and resources and district-level support and expectations have a direct impact on the principal. Thomas Guskey's research (2000) indicates that an emphasis on high-quality, ongoing professional development can not only change organizational patterns and norms but can lead to improved student outcomes. Figure 2.1 provides the theoretical framework for this improvement project.

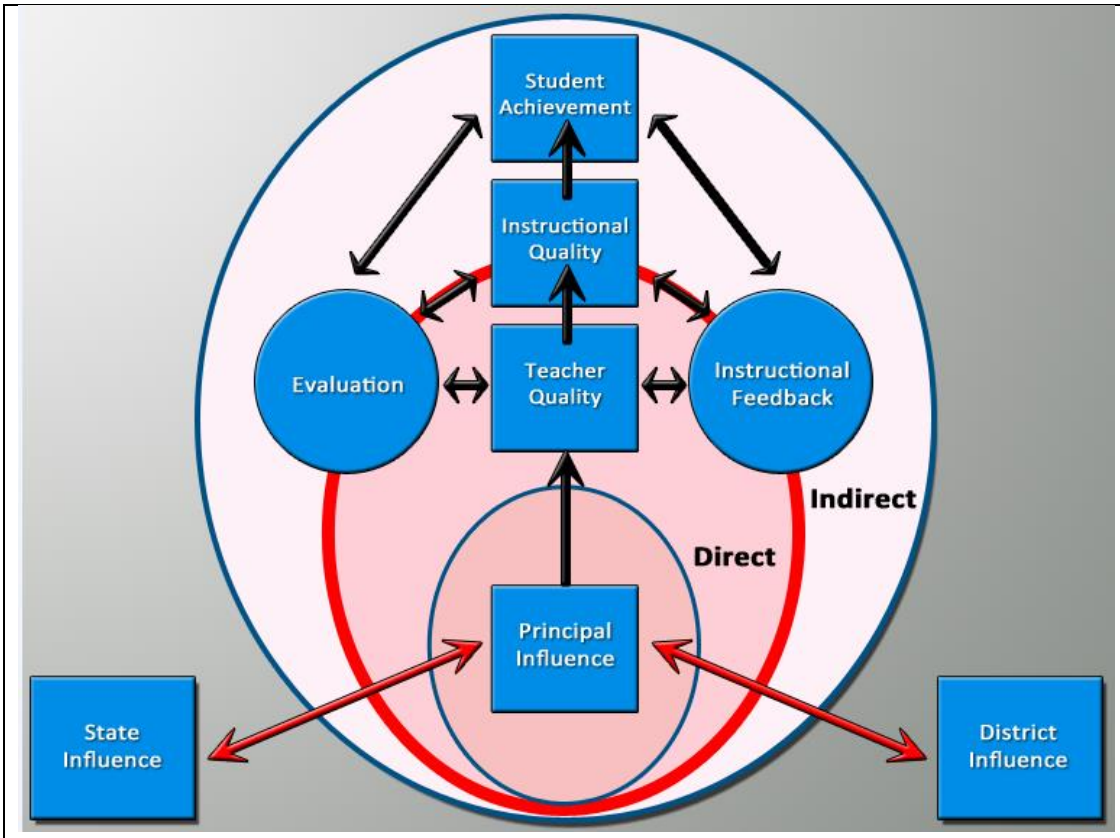


Figure 2.1. An Adaptation of *The Ripple Effect*

Figure 2.1 is an adaptation of "The Ripple Effect" design developed by Clifford, Behrstock-Sherratt, and Fetters (2012). In the original design, the two factors that affect principal quality are "district and community contexts" and "school conditions."

If both the state and district-level leaders work together to strategically and purposefully build learning and growth opportunities for principals, the adaptation of "The Ripple Effect" will yield not only principals that are more skilled evaluators but through the concept of the ripples, they will, in fact, also support an improvement in student achievement through the indirect effect they have through accurate evaluation, instructional feedback and coaching they provide to teachers (Simmons & Mullins, 2013).

Desired State

District leaders in Concordia want to identify and implement systems and processes that will promote deep change in evaluation practices throughout the school district. The goal is that teacher evaluation data will parallel educator value-added. If principals rate teachers accurately, CPS should not only see greater correlation between value-added ratings and principal ratings on summative evaluations but also should see improvement in high-stakes assessment proficiency and growth among students. Figure 2.2 provides an implementation framework for how the NCDPI and CPS worked jointly to develop a plan to provide prolonged engagement around the OCT Pilot through the specific targeted strategies intended to enhance the content through collaborative viewing and discussion sessions with a focus on the language of the standards.

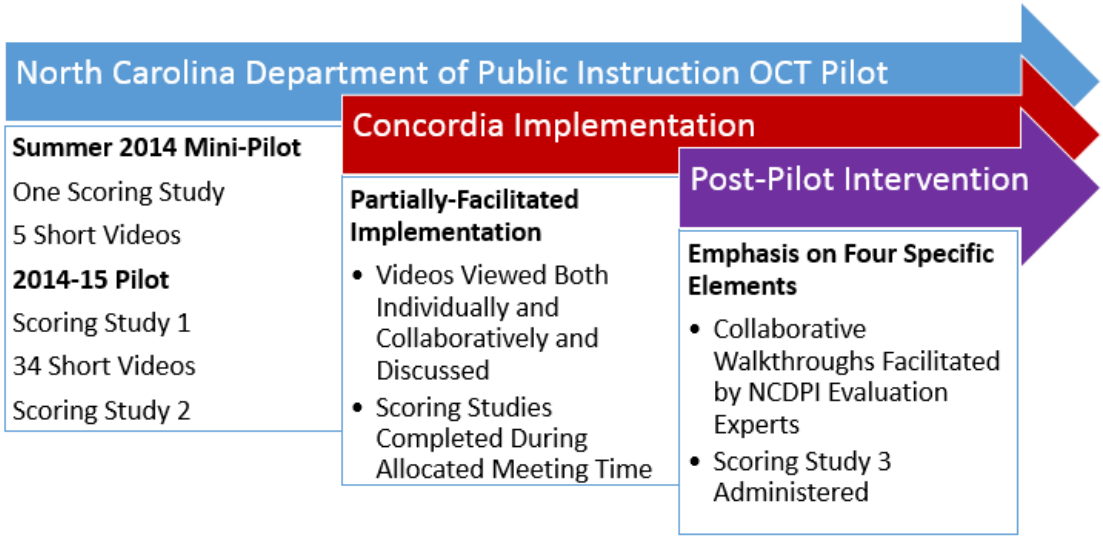


Figure 2.2. Phases of OCT Pilot implementation for Concordia Public Schools

Previous Initiatives

Concordia Public Schools has, in the past, provided some support for the North Carolina evaluation for principals. Prior to the 2013-14 school year, the director of human resources provided the majority of this support. However, beginning with the 2013-14 school year, the curriculum team began to supplement that support by ensuring that evaluation professional development and training had a more concerted emphasis on curriculum and instruction. The superintendent shifted the oversight of Title II, which includes professional development, from the director of human resources to the chief academic officer. In a review of the Title II plan, the curriculum team developed an intentional plan to address curriculum and instruction through evaluation practices.

Table 2.2

Previous Initiatives to Support Concordia Principals with the NCEES

Year	Event/Training	Facilitator(s)
2007-2008	CPS selected to pilot the new North Carolina Educator Evaluation System	N/A
2007-2008	Introductory training to the new North Carolina Educator Evaluation process	Director of Human Resources
	Major change emphasized was instrument no longer measured a "snapshot" in time but a longitudinal, comprehensive look at a teacher's performance over the course of the academic year	
2007-2008	Principals and APs trained on NCEES together	Director of Human Resources
2008-2009	Principals, APs, and curriculum directors develop a "look-fors" sheet for the 25 elements embedded within the five teacher evaluation standards	Director of Human Resources
2008-2009	All teachers and certified staff trained on the NCEES prior to the start of the 2008-2009 year.	Director of Human Resources
2008-2009	Follow up session with principals after the first nine weeks for questions and concerns related to the new tool and new processes	Director of Human Resources
2008-2009	Individual conferences with principals to address specific questions/concerns related to the tool and process	Director of Human Resources
Summer 2009	Revisited CPS "look-fors" document to make revisions and updates	Director of Human Resources
2009-2013	Director of Human Resources, selected principals, and curriculum directors attended multiple NCDPI sponsored Regional Education Service Alliance (RESA) professional development sessions related to the NCEES process or tool	NCDPI Educator Effectiveness Professional Development Consultants
2009-present	All certified new hires trained yearly (either beginning of year or beginning of second semester) on NCEES process and tool	Director of Human Resources

Recent Initiatives

Over the course of the 2013-14 and 2014-15 school years, district leaders initiated a variety of interventions to address the problem. The goals of the initial interventions were to provide common frameworks, structures, and expectations central to evaluation in Concordia Public Schools. Each of the following interventions was designed to support improvement in the quality of evaluation throughout the district.

Collaborative Walkthroughs

Concordia principals engaged in two collaborative walkthroughs in teams of four during monthly principals' meetings. Central office personnel visited each school weekly to conduct collaborative walkthroughs with a principal or assistant principal to improve target agreement and to gauge instructional practices in their schools. These walkthroughs were focused around the following instructional expectations:

- Utilizing technology to improve instruction
- Implementing opportunities for students to engage in the 4Cs – collaboration, communication, critical thinking, creativity
- Engaging students
- Differentiating instruction

All principals had an opportunity to provide input on the walkthrough document in principals' meetings before the tool was finalized during the 2013-14 school year. The tool was modified on August 6, 2014 for the 2014-15 school year based on results from the 2013-14 results, conversations, and principal feedback. Principals, assistant principals, and district leaders committed to completing five classroom walkthroughs each week. Furthermore, at monthly principals' meetings, district leaders and principals

engaged in small group collaborative walkthroughs and engaged in small group discussions following each walkthrough.

Revelations from walkthroughs. Based on discussions in principals' meetings, overall understanding of the language levels of engagement, and agreement during classroom walkthroughs, the collaborative walkthrough process has led to a deeper understanding of the concept of engagement. Through evidences captured in the walkthrough document, the principals and district leaders have learned that although the use of essential questions and clear learning targets is a district expectation, levels of implementation vary from school to school. Based on the quality of the learning targets and essential questions posted, one of the most important realizations is that teachers needed more training on developing learning goals for their students. Moreover, walkthrough data suggested teachers needed additional training on understanding the difference between clear learning targets and essential questions as well as how to communicate these goals to students.

Professional development needs that emerged. Principals received professional development on the Technological, Pedagogical, and Content Knowledge (TPACK) model in fall 2013 and the Substitution, Augmentation, Modification, Redefinition (SAMR) model of technology integration in September 2014 after they expressed a need for support to help them understand SAMR and how to provide feedback to teachers. Principals also learned how to recognize whether or not teachers were implementing TPACK when using instructional technology and how to recognize differing levels of the SAMR model of technology integration. Ongoing collaborative discussions during walkthroughs and whole group discussions following walkthroughs provided district

leaders with ongoing formative feedback as to how well principals could pinpoint elements of TPACK and SAMR during walkthroughs.

Principal perception of walkthroughs. Principals expressed that they enjoyed the discourse following each of the walkthroughs, and they reviewed the data from both the district and their own schools to make decisions for future professional development. During principals' meetings, the principals discussed that talking about instruction improved rater-agreement in that they were able to listen to others share their reflections on each element of the classroom walkthrough tool. Moreover, they remarked that discussing what others saw in classrooms and how each observer approached the walkthrough was helpful in improving their own evaluation practices.

Analysis of NCEES Standards 1 and 4

On July 7-8, 2014 and August 7, 2014 during the NCDPI-sponsored Summer Institute and Concordia Summer Leadership Retreat, principals, assistant principals, instructional coaches, selected teachers, and district leaders worked together to analyze all elements of standards 1 and 4 of the North Carolina Professional Teaching Standards to gain a deeper understanding of each element and to improve both rating accuracy and target agreement. The exercise consisted of discussing the verbiage of each element and of each rating for the particular element. As a group, participants developed "look-fors" in order to improve rater-agreement. The "look-fors" were provided to each principal for review and for use during the 2014-15 school year. Standards 1 and 4 were selected as the focal standards because all teachers are evaluated on these two standards every year. Only teachers in their license renewal year are evaluated on all five standards.

Evaluation Case Study

On August 6, 2014, Concordia principals received training and support regarding how to develop appropriate documentation through a two-part evaluation case study where they had to work in teams to provide formal documentation based on a teacher's evaluations and ongoing performance. This training served as the logical next step for principals who had been focused on "look-fors" and rater-agreement but had not demonstrated that they had applied this knowledge of the standards to their summative evaluations. Based on the instructional conversations regarding ratings during collaborative walkthrough conversations, it is evident that principals in CPS have high instructional expectations. However, summative evaluation results from the 2013-2014 school year revealed that only one teacher in CPS who taught a tested subject was rated developing on any standard, and that teacher was rated "developing" on only one standard. This teacher received a rating of "meets expected growth" on standard 6. This discrepancy revealed a need to offer support beyond that previously provided to ensure accurate ratings. District leaders agreed that by providing support with documentation and feedback to principals that all site-based administrators would have the appropriate tools, such as sample letters and sample action steps to provide documentation, when appropriate, regarding instructional practice. Training materials and resources are housed online for CPS internal use.

Developing Improvement Plans

On November 10, 2014, Concordia's Board attorney provided professional development for principals on how to develop high-quality, legally-compliant mandatory improvement plans, monitored improvement plans, and directed improvement plans. The

need for this training arose when district leaders discovered that the principals had been misinformed about the appropriate documents to use when developing improvement plans, when and how to place documentation in a teacher's personnel folder, and how to ensure legal compliance when working with marginal teachers. Providing principals with important legal information regarding evaluation and documentation was essential.

After the training provided by the Concordia Board attorney, four improvement plans were written during the 2014-2015 school year. All four plans were developed using the proper forms and added to the NCEES system appropriately. Principals are using the training to improve their practice and are now in compliance with state law.

CHAPTER 3: INTERVENTION DESIGN

The lack of correlation between the ratings principals assign teachers and teachers' value-added scores indicates that principals lack the requisite knowledge and skills to evaluate teachers and need further training. Student performance on state-mandated, standardized assessments at CPS demonstrated that the district needed to develop a comprehensive plan to support administrators who provided guidance to teachers in the areas of curriculum implementation and instructional practice. This problem is a developmental one most closely aligned with the need to build capacity in the site-based administration. The first step in solving this problem was sharing three types of data with principals: evaluation data, student proficiency data, and student/teacher growth data. The next step was to allow them grapple with our current reality. Next, it was important to review past interventions and principal perceptions to design and provide meaningful professional development supports that would engender trust from principals and have the best chance for success.

Kruse and Louis (2009) contend that the key to understanding a school or organization's culture is to review and understand the following conditions: professional community, organizational learning, and trust (PCOLT). Concordia's leadership worked with site-based administrators to build a strong sense of professional community and organizational learning through several means. Bi-monthly principals' meetings ensured district and site-based leaders had structured opportunities to engage in discourse, learn together, review data, and set goals. Trust was cultivated through a purposeful development of a culture of teamwork. The group celebrated together and built time into

each academic year for team building and socializing. When organizations "foster their ability to create community, learn together and engender trust in each other, improvements in the outcomes for students are improved" (Kruse & Louis, 2009, page 14). The PCOLT model is embraced in Concordia, and district leadership preserved time and resources to ensure that professional community, organizational learning, and trust were protected.

With the aim of improving principal expertise in evaluating teachers as the focal point for support for principals during the 2014-15 school year, the intervention design was developed around the idea of collaboration through ongoing and consistent support. Based on the feedback from the collaborative walkthroughs during the previous school year as well as feedback from participation in the NCDPI OCT Mini-Pilot, the design team became keenly aware of the importance of providing opportunities for collaboration in an open, trusting environment where principals and central office leaders would learn together.

Intervention Design Plan

Providing professional development experiences for administrators where they can learn collaboratively by reflecting on their evaluation experiences and discuss evaluation practices is an important part of the learning process. A comprehensive study by Yoon et al. (2007) suggests that the duration and the sustained, ongoing nature of professional development plays an important role in the success of the initiative. Furthermore, the study indicates that both continuous follow up and on-going, job-embedded opportunities for discussion, feedback, and continued emphasis are all crucial elements of implementation success. District leaders can ensure that evaluators have

opportunities to reflect on their practice, specifically in terms of professional development implementation, by providing focused, on-going professional development as part of their regularly scheduled meetings or in sessions designed specifically around reflection, sharing, and feedback.

The evaluation process is important on many levels. Evaluators provide teachers with clear, specific feedback to help them grow and improve as teachers. However, before feedback could be addressed, the district needed evidence that evaluators had a deep understanding of the North Carolina Educator Evaluation Process, the North Carolina Professional Teaching Standards and how to rate those standards during a classroom observation. The goal of this improvement project was to provide opportunities for principals to engage in ongoing discourse around evaluation, the North Carolina Professional Teaching Standards, and the process of gaining a deep understanding of the Rubric for Evaluating North Carolina Teachers through a series of improvement strategies.

The original implementation plan proposal was presented in February 2014. This plan proposed a 90-day cycle of interventions beginning in March 2015. However, during implementation, data prompted that decisions be made to abandon some aspects of the original design while continuing to embrace others. Figure 3.1 provides a framework for the proposed interventions.

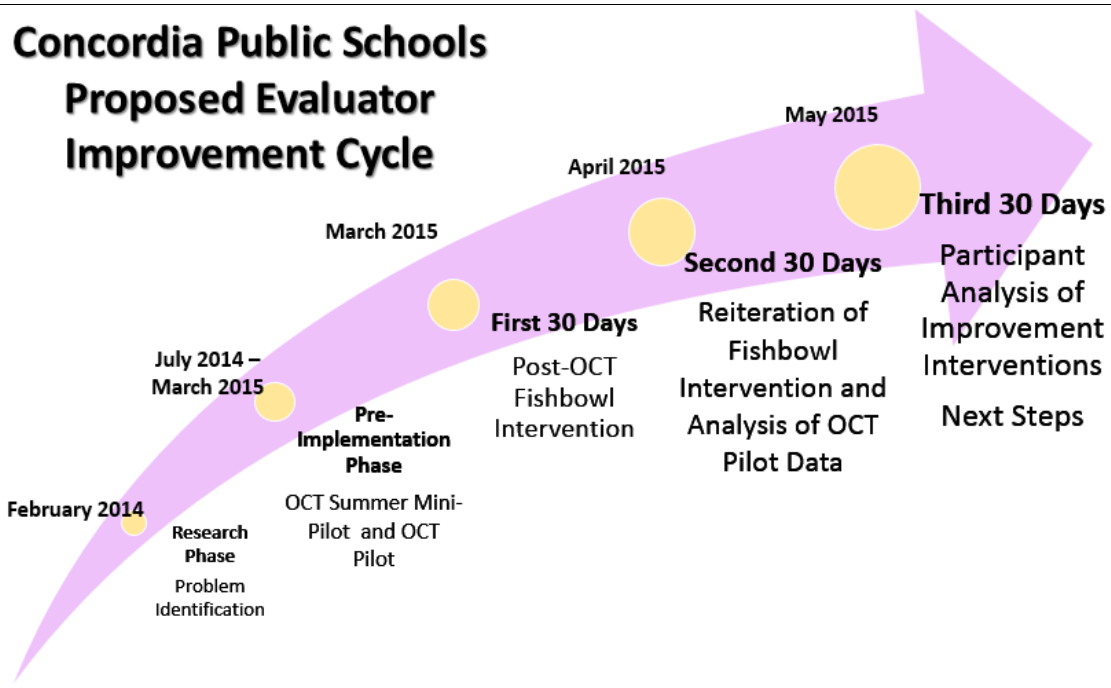


Figure 3.1. Concordia Public Schools proposed evaluation improvement cycle

Proposed Implementation Plan Overview

Originally, the improvement project was slated to begin after Concordia's evaluators participated in two phases of the NCDPI OCT Pilot. The first phase of the pilot was a mini-pilot the state offered during July 2014. During this OCT Mini-Pilot, participants reviewed classroom lessons via online video platform, rated those teachers using the Rubric for Evaluating North Carolina Teachers, and received feedback on their ratings. Following the OCT Pilot, the plan was for Concordia evaluators to take part in the full OCT Pilot from November 2014 – March 2015. Interventions were to begin after the OCT Pilot. However, participation in the OCT Mini-Pilot and feedback from participants led the design team to reevaluate the original plan, abandon the March 2015 start, and opt for incorporating the OCT Pilot as a viable first intervention.

Proposed Implementation Plan - First 30 days

The original implementation plan was for a 90-day cycle to be designed around rater-agreement and accuracy while conducting collaborative abbreviated observations (20 minutes). The first intervention was slated to take place in April 2015 and was to consist of Concordia principals accompanying an expert from the Educator Effectiveness Division of the North Carolina Department of Public Instruction into classrooms in groups of three or four to conduct abbreviated observations. Two groups would simultaneously observe two classrooms. Principals would be asked to rate teachers on specific observable elements. After the observations, participants would come together and take part in a "fishbowl" discussion with an inner and outer circle. One group would engage in dialogue in the inner circle. The plan for this phase of implementation was to have an NCDPI field expert serve as the facilitator for this discussion. The principals would discuss ratings, "look-fors", specific classroom examples, and questions they had about the observation. Meanwhile, the remaining principals and curriculum directors would sit in an outer circle around the principals and take notes on what they heard, questions they had, insights they gleaned from the discussion, and other important moments of confusion or clarity. After the discussion, groups would then switch positions, and the group that was participating in the discussion would then become the note takers while the note takers would become the participants in the discussion. All principal and director notes were to be collected and coded. The session was designed to conclude with an opportunity for principals to reflect on what they gleaned from the exercise and what they needed next.

Proposed Implementation Plan - Second 30 days

The original intervention plan was abandoned during the implementation. However, the second iteration of the project was originally designed to capitalize on the data collected during the first 30 days. Qualitative data from the notes taken in the fishbowl activity were to be coded and analyzed to determine individual and group needs. Furthermore, I anticipated that OCT Pilot data would be available for review based on the information I had been provided by the NCEES Consultant. The initial plan was to review data from the two scoring studies that served as a pre-assessment and post-assessment during the NCDPI-initiated OCT Pilot during the fall and spring of the 2014-2015 school year as well as the qualitative data gathered from the fishbowl discussions, and based on those reviews determine needs and develop the next iteration of the project. The design team was to review all data before making recommendations to the implementation team. The second intervention was slated to take place in April 2015. This cycle became the third phase of the project due to availability of data from the OCT Pilot as well as limitations of time. Both of these variables led the design team to make modifications to the initial proposal.

Proposed Implementation Plan - Third 30 days

The original proposal included a third 30-day cycle that later became repurposed. Although it was not completed as an independent cycle, this cycle was embedded into the final implementation plan. Originally, the design team planned to have principals engage in a deep reflection of each of the interventions implemented to improve evaluation and share their thoughts regarding which interventions were most effective. They were also going to engage in a review of their own evaluations of teachers in their buildings using

modified versions of the National School Reform Faculty protocols *Chalk Talk* (Appendix G) and *Critical Friends Work Groups* (Appendix H). The goal of this cycle was to use these reflections to provide district leaders with data and information to plan for next steps of support with evaluation. Furthermore, the design and implementation teams believed principals would be empowered to identify their own areas of strengths and improvement and create their own next steps as professional learning goals for the 2015-16 school year. District leadership planned to support the principals and to use these goals to inform trainings and professional learning opportunities.

Aim

The aim of this improvement project was to improve the quality of evaluation by improving target agreement and accuracy and reducing scoring bias by implementing a variety of improvement strategies designed with the support and guidance from the NCDPI NCEES Consultant. These strategies included a partially-facilitated model of the North Carolina OCT Pilot, collaborative classroom walkthroughs with facilitated Socratic-style discussions, and the collection of perception feedback on the quality of the interventions as well as self-reported needs. The interventions that occurred prior to the Phase I cycle and the outcomes of these interventions informed Phase I and subsequent phases of the project.

Expected Outcomes

Expected outcomes were presented in goals. The goals for this project include:

- Concordia Public Schools evaluators and district leaders participated in the 2014-15 OCT Pilot with the expectation that target agreement would increase by March 2015, as measured by a comparison of Scoring Study 1 and Scoring Study 2.

- Concordia Public Schools evaluators and district leaders participated in the 2014-15 OCT Pilot with the expectation that scoring bias would decrease by March 2015, as measured by a comparison of Scoring Study 1 and Scoring Study 2.
- Concordia Public Schools evaluators and district leaders participated in the 2014-15 OCT Pilot with the expectation that rater discrepancy would decrease by March 2015, as measured by a comparison of Scoring Study 1 and Scoring Study 2.
- Concordia Public Schools principals participated in the 2014-15 OCT Pilot and subsequent interventions with the expectation that the correlation between principal ratings on Standards 1-5 of the NCEES Rubric and Standard 6 would increase by October 2015, as measured by a comparison of EVAAS data from the 2013-14 school year and the 2014-15 school year.

Design and Implementation Teams

Design team. As chief academic officer in CPS, I organized and led the design team. Other members of the team included the NCDPI NCEES Consultant, the director of elementary education, the director of high schools, the director of exceptional children and the principal of Central Concordia Elementary School. During the Pre-Cycle phase, the design team met to examine evaluation data, assessment data, and value-added data. They also reviewed the evaluation-related interventions that had been implemented since the 2013-14 school year. Throughout the study, the design team held formal and informal meetings to analyze data, develop protocols, design next steps, and consult with the implementation team.

Implementation team. The implementation team brought a variety of experience to the project. As chief academic officer and leader of the project, I brought experience as a professional development consultant for the NCDPI and was a NCEES trainer. Other team members brought similar expertise. The NCDPI NCEES Consultant was also a member of our implementation team. The main job function for the NCDPI NCEES Consultant is to implement an effective evaluation system for the state. The NCEES Consultant was directly responsible for the development of the OCT Pilot and has worked with Concordia's principals both online and in face-to-face sessions. She conducts bi-annual meetings for principals in each of the eight regions of the state to share updates, provides professional development support for evaluation, and supports principals by allowing them to provide feedback on many evaluation-related issues. The director of elementary education has served in Concordia Public Schools as a teacher, assistant principal, principal, and director of elementary education. She has a strong understanding of evaluation and works closely with elementary principals and instructional coaches to provide support with evaluation, feedback, and monitoring. The director of exceptional children began her career as an elementary school teacher who then moved to an instructional coach position. She served as an assistant principal and was a principal for more than four years. She was named the regional principal of the year as well. The director of accountability and technology primarily served in the capacity of ensuring technical needs were addressed and that connectivity, the online platform, the log-in site for the OCT Pilot, and other technical aspects of the intervention were easily accessible for all participants.

Methods for Periodic Assessment of Intervention

The state OCT Pilot yielded three quantitative measures - rater accuracy, rater agreement, and scoring bias. However, at the micro-level, a more personalized and specific means of gathering data to determine participants' needs and areas for growth was necessary. Qualitative data was collected from interviews with human resource directors, anecdotal notes from principals' meetings, reflections and insights participants shared on tools used during interventions, and the videotaped "Fishbowl" Intervention. Data were merged by reporting statistical quantitative data through tables, graphs, and charts and incorporating and providing qualitative statements, ideas and themes that either supported or refuted the results of quantitative analysis. This practice, as described by Sandelowski, Volis, & Knafl, (2009) drove the integration process.

Instrument Validation

Several instruments were used to collect data over the course of the research phases. The OCT Pilot scoring studies and lessons were used to collect most of the quantitative data. Perception data, reflection data, and anecdotal data were collected throughout the study through notes, reflective handouts, and videotaped discussions. Interviews were conducted with human resource directors and data were collected from NCDPI Principal READY meetings in an attempt to collect perception data from sources outside the district as well.

OCT Pilot validation. The OCT Pilot included two scoring studies that served as pre-test and post-test. The scoring studies each consisted of a full-length (approximately 45 minutes) video that was on all 17 observable elements from the Rubric for Evaluating North Carolina Teachers. "The chief way we collect content-related evidence of validity

is judgmental, that is, by asking competent individuals to scrutinize the items on a test to judge whether whatever the test is measuring has been satisfactorily represented." (Popham, 2010, p. 24). In addition to the NCDPI Educator Evaluation Team, six expert scorers worked under direction of psychometrician, Dr. Joshua Priddy, formerly of Mid-continent Research for Education and Learning International (McREL), who developed rigorous protocols for determining master ratings. McREL International is a non-profit that provides scientifically-based research and evaluation services to educational agencies of all levels. Priddy's team, in conjunction with the NCDPI Evaluation Team, worked together and reviewed and rated each video numerous times to ensure accuracy. According to Simmons, the NCDPI review team consisted of six subject matter experts with backgrounds in education as instructors, principals, and educational methodology trainers (personal communication, August 13, 2014). Some members of the review team also served on the original team that helped develop the Rubric for Evaluating North Carolina Teachers.

The OCT Pilot consisted of a two scoring studies that were administered to participants as a pre-assessment and post-assessment. Between these two assessments, participants viewed 34 short excerpts of classroom videos. Prior to the pilot, the NCDPI review team had reviewed and rated these 34 short videos. Each of the short videos was approximately three to five minutes in length and was designed to give participants an opportunity to focus on one specific element of the Rubric for Evaluating North Carolina Teachers. Participants viewed and rated two videos for each of the 17 observable elements in the rubric. The team, led by Priddy, also rated two additional full-length

videos that were housed in the online OCT platform. These videos were not assigned to participants in the OCT Pilot but were available on the Bloomboard online platform.

The same set of master raters developed the ratings for all videos in the OCT Pilot, the scoring studies and the lessons. Each of the raters had expertise and experience with the Rubric for Evaluating North Carolina Teachers at various levels, and the process was led by and validated by a psychometrician working for McREL, a respected national leader in scientifically-based educational research methods. Furthermore, the NCDPI has approved the OCT Pilot ratings for research across the state.

The NCDPI also demonstrated what was perceived to be a commitment to continuous quality improvement as they asked for specific feedback on the rating of each item and collected that feedback both in the online platform in an open-ended text box and in through feedback sessions during the OCT Mini-Pilot. These opportunities to collect qualitative feedback on the platform, the OCT, and the evidences and ratings will result in improvement if the state uses the feedback to reevaluate the ratings.

OCT Pilot platform. NCDPI partnered with BloomBoard and Empirical Education to house the Observation Calibration Training platform and analyze findings from the NCDPI study to present results and findings. Bloomboard is a California-based educational development company that provides personalized professional development. BloomBoard was selected from six vendors to serve as the vehicle to house and deliver a fully functional system for online professional development. Empirical Education is a research company that supports evidence-based decision making for school districts through the use of tools and services. Empirical Education served as the vendor who analyzed the data from the OCT Mini-Pilot and OCT Pilot to provide the NCDPI with

quantitative results regarding the impact of the study. The design team in Concordia contacted representatives from NCDPI, Bloomboard, and Empirical Education to receive raw data as well as case study data from the companies' analyses for use in this project.

Data Analysis

Although mixed methods research is challenging in terms of combining data, the conceptual design of this improvement project inherently lent itself to employ mixed-methods design. Data analysis took place in overlapping phases: review of instrument validation, quantitative data analysis, qualitative data analysis, and ongoing decision-making following data analysis phases. Table 3.1 provides an overview of data collection methods as well as the data analysis strategy for each data set.

Table 3.1

Concordia Evaluation Improvement Project Data Analysis Matrix

Improvement Effort	**Evidence of Improvement Collection Strategy**	**Frequency / Threshold for Intervention Modification**	**Analysis Strategy**
Improve Rater Accuracy through the OCT Pilot	Comparison of First, Second, and Third Scoring Studies in the OCT Pilot	November 2014, April 2015, and May 2015	Paired Samples t-test
	Comparison of EVAAS Data with Evaluator-Rated Standards	October 2013, October 2014, October 2015	Spearman-R Correlation
District-Wide OCT Collaborative Rating Discussions with Evaluators	Collaborative note-taking by implementation team	Bi-monthly January 2015 – March 2015	In Vivo Coding
Collaborative Walkthroughs/ Fishbowl Activity	Videotaped Fishbowl Activity "Capture Your Thoughts" Tool Participant Reflection	April 2015	In Vivo Coding and Pattern Coding

Comparison of OCT scoring studies. The Observation Calibration Training Pilot included two scoring studies, Scoring Study 1 (SS1) that served as a pre-test and Scoring Study 2 (SS2) that served as a post-test. The scoring studies each consisted of an approximately 45-minute video of a teacher teaching in a classroom. The videos were acquired by the NCDPI from the Measures of Effective Teaching (MET) library from the

Bill and Melinda Gates Foundation. Participants rated teachers in these scoring study videos on all 17 observable elements on the Rubric for Evaluating North Carolina Teachers. A paired samples t-test was used to compare the means of a normally distributed interval dependent variable for two independent groups. The same 12 participants completed both scoring studies which ensured the measure was reliable. This is a repeated measures design.

The second intervention in the project included Scoring Study 3 (SS3) as an additional assessment of the impact of the second phase of interventions. SS3 also consisted of a full-length classroom video in which participants rated all 17 observable elements. Again, a paired samples t-test was used to compare participants' performance. Due to the nature of the second phase of the project, the paired-samples t-test was only conducted on the four elements that served as a focus for the intervention, and the comparison was made between the participants' performance on SS2 and SS3.

Correlation of EVAAS data. All seven principals in Concordia Public Schools served in the same post during the 2014-15 school year as they served during the 2013-14 school year. A Spearman-R correlation was used to measure how ratings evaluators assigned teachers on observable standards correlated to the standard that was populated from value-added data through EVAAS for the 2013-14 school year. The same Spearman-R correlation was used to determine correlation during the 2014-15 school year. McDonald (2014) suggests that researchers use Spearman-R to determine whether two ranked variables covary. The design team analyzed the data from the 2013-2014 and 2014-2015 school years to determine the impact of the interventions on evaluator practice.

Collaborative rating discussions. OCT Pilot participants in Concordia Public Schools engaged in bi-monthly rating discussions after viewing OCT videos on one particular observable element during principals' meetings. The implementation team engaged by taking anecdotal notes during some discussions. The team received instructions to capture exact words and phrases as much as possible. Furthermore, following the discussions, the implementation team documented their reflections of the discussions. The team did not elect to record the discussions in an effort to ensure participants felt comfortable voicing their confusion and misconceptions. In vivo coding was used to honor the participants' voices in order to "stay as close as possible to research participants' own words or use their own terms because they capture a key element of what is being described" (King, 2008, p. 473-474). Codes were analyzed to determine participant needs and areas for improvement in order to develop systems of support with evaluation.

Fishbowl discussion. During the Phase II Fishbowl Intervention, participants were divided into two teams to conduct a collaborative, 20-minute observation of a teacher at Southwest Elementary School. Following the observation, groups engaged in a Socratic-style "fishbowl" discussion led by an evaluation expert from the Educator Effectiveness Division at the NCDPI. During a "fishbowl" discussion, half of the group sits in a circle and engages in a facilitated discussion around a topic. The remaining half of the group creates an outer circle around the group engaging in facilitated discussion and uses a process to capture insights, thoughts, and questions to share at the conclusion of the discussion. After the discussion ended, the groups switched positions, and the groups engaged in the process again.

The Fishbowl Intervention was videotaped, and the transcript was coded using in vivo coding as well as pattern coding. By employing both methods, the transcript was analyzed for exact words and phrases as well as for categories that emerged as pertinent to improving evaluation. Participants on the outside of the fishbowl completed a "Capture your Thoughts" handout as a means to preserve their thoughts about the conversation, share insights and questions, and make connections to their own practice. (See Appendix I).

Changes to the Implementation Design

When implementing improvement science projects, often what is planned as particular interventions or stages is modified because unlike conducting research on institutions and data outside of one's sphere of influence, when real-time modifications can be made to improve the quality of the interventions or better meet the needs of participants, the design team can use data to justify in-the-moment adaptations to the original plan. In this project, several specific occurrences led to changes in the implementation plan.

OCT Mini-Pilot Data

Prior to participation in the OCT Mini-Pilot, I was unaware that we would receive raw data from this activity. My original goal for participating in the pilot was to give evaluators in Concordia an idea about what to expect in the OCT Pilot and to give them a jump-start on navigating the OCT online platform. However, the data we received from the one full-length scoring study and five short video lessons in the mini-pilot as well as the feedback our participants provided to both NCDPI and Bloomboard actually helped drive the next phases of the project.

OCT Pilot Data

When discussing the OCT Pilot with the NCEES Consultant, I was aware only that we would receive data from both scoring studies. I did not realize we would also be provided with data for the 34 individual short videos or additional data regarding scoring bias and rater discrepancy. The scale of data our design and implementation teams had for review led us to a much deeper analysis of this phase of the project and provided much more of an impact than we had anticipated.

When the team reviewed the OCT Pilot data and realized the scope and implications of those data, important changes to the original plan were made to meet the needs indicated by the data. When practitioners implement improvement science with fidelity, they must look beyond the theoretical scope of empirical research and use available data in real-time to make meaningful, potentially transformational decisions for their institutions.

CHAPTER 4: IMPLEMENTATION AND RESULTS

As with any practitioner-initiated plan, circumstances and available information have an undeniable effect on implementation. The intervention plan that was presented in February 2015 changed due to a number of circumstances and data available. Fortunately, more data was available than originally anticipated, and the access to that data was provided in a more timely manner than originally suggested. Therefore, data-driven decisions were made to modify the design and implementation plans to provide a deeper, richer, timelier experience for the participants from CPS.

In this chapter, I will discuss the ways in which the original intervention plan was modified and why and present evidence to illustrate that the quantitative results from the pilot and the intervention were clear. Data collected after each intervention suggested that the OCT Pilot intervention did improve target agreement and decrease both scoring bias and rater discrepancy. Data also suggest that the Fishbowl intervention also had a positive impact on target agreement. Additionally, the qualitative data collected from discussions, informal meetings, and surveys suggest that not only were the interventions successful but also that participants both enjoyed and appreciated the opportunity to engage in guided discourse around the standards and evaluation. This chapter provides a detailed description of each phase and how the data and outcomes led to the next iteration of improvement.

Modified Implementation Plan

The original intervention plan presented in February 2015 was modified for several reasons. All changes to the original plan were discussed by the design and

implementation teams and deemed best for the participants. The rationale for each of the changes was based on time, data, or other circumstances out of our control. Although the intervention plan and the methods used in the project did change as data became available and circumstances changed, the goals that drove the project did not change.

Rationale for Selected Interventions

OCT Mini-Pilot data. One of the main reasons for modifying the original implementation plan was the availability of data from both the OCT Mini-Pilot and OCT Pilot. This data helped to drive the decisions of the design and implementation teams. First, the state-initiated OCT Pilot began in late November 2014. Originally, I believed that the OCT Pilot would provide the team the baseline data we needed to actually begin interventions. However, we were provided the raw data from Empirical Education from the summer 2014 OCT Mini-Pilot (n=7) that gave the team an indication of our district's current reality in terms of target agreement. All seven participants in the summer 2014 OCT Mini-Pilot were also leaders who participated in the OCT Pilot and the subsequent interventions. The availability of both the OCT Mini-Pilot data and the comments and opinions of the seven participants in the OCT Mini-Pilot led us to make modifications to the plan.

The qualitative results from the OCT Mini-Pilot Scoring Study indicated that participants in CPS agreed with target ratings only 44% of time. Average percent discrepant was 12%, and 43% of participants exhibited scoring bias. These data indicate CPS evaluators need support with rating for accuracy, target agreement, and scoring bias. Furthermore, after participation in the OCT Mini-Pilot, Concordia principals were eager to participate in the full OCT Pilot during the 2014-15 school year. They expressed a

need for more work with target agreement and shared that they appreciated the opportunity to talk about instruction both in informal conversations at principals' meetings and during the NCDPI feedback sessions on July 11-12, 2014.

OCT Pilot data. The most important data we received that provided us with greater insight was the ongoing OCT Data from SS1, the 17 modules, and SS2. This data was available sooner than we had anticipated, and we were able to map the group's growth or lack thereof by using multiple data points, including the OCT Mini-Pilot, SS1, the 17 modules, and SS2. I worked with the curriculum team to review this data before discussing potential modifications to the pilot. However, the vast amount of data indicated that the OCT itself had a statistically significant impact on scoring bias and target agreement. Therefore, we realized that the OCT Pilot did not supply our baseline data, SS1 did. Therefore, the OCT Pilot became our first intervention.

The OCT Pilot consisted of SS1, a full-length class video where participants rated all seventeen observable elements of the North Carolina Professional Teaching Standards (NCPTS); 34 short videos from two minutes to six minutes where participants rated only one of the seventeen observable elements; and SS2, another full-length class video where participants again rated all seventeen observable elements on the NCEES rubric.

The OCT Pilot provided principals and central office leaders with the opportunity to evaluate and rate instruction. NCDPI provided loose guidelines for OCT participation. However, one of the elements participants in the OCT Mini-Pilot indicated was important was the ability to collaborate and discuss the standards, the wording of the standards, and how to interpret the "look-fors" without subjectivity. Therefore, the implementation team in Concordia Public Schools developed more structured and supported guidelines as

well as a more interactive and rigorous process for OCT Pilot implementation that included opportunities to view some of the OCT module videos together to discuss ratings, "look-fors," and questions after each participant submitted his/her rating in the BloomBoard platform.

Another important reason for choosing the OCT Pilot as an appropriate intervention was pilot alignment with need. One purpose of the pilot was to address the state-wide issue of lack of correlation between principal ratings and value-added scores, the exact need in CPS. The pilot study would also yield important pre-test and post-test qualitative results that would measure whether or not the OCT Pilot may have had an impact on principal practice.

Scoring Study 3. Once I realized that the SS1 could serve as our pre-test and SS2 as our post-test, it made perfect sense that, whether it was planned or not, the OCT was an intervention. Although this five-month pilot did not fit into the 90-day cycle that I had planned, it was a much more purposeful intervention that yielded the data we needed to move forward. Once the design team realized we would be making a vast shift in our intervention plan, we contacted NCDPI to discuss options for further assessments. If SS1 and SS2 were specifically designed to measure progress and they had been psychometrically validated, then we were in need of another scoring study to be able to re-assess participant progress after further interventions. The NCDPI NCEES Consultant spoke with the Director of Educator Effectiveness and with both Bloomboard and Empirical Education, and they released SS3 to us to use as another assessment following our post-OCT intervention. SS3 had been scored and psychometrically validated in the same manner and at the same time as SS1 and SS2. A paired samples t-test was used to

compare the means, and the same 12 participants completed SS3. Results are displayed later in this chapter.

Modified Design Plan

The intervention framework shifted from a 90-day cycle framework to five Plan-Do-Study-Act (PDSA) cycles that loosely fit into the 90-day cycle framework. A PDSA cycle is a repeating problem-solving process used to improve a process or implement a change. One iteration of a change or modification is introduced and tested. Results are studied, and leaders use the data to inform the next cycle in which a modification or change is added to address the need identified in the cycle. This modification was developed and implemented by the design and implementation teams along with input from the superintendent and intervention participants. The Research Phase and Pre-Implementation Phase provided more local data and impetus for the three intervention phases in Concordia Public Schools prior to the interventions studied during the course of this project. The Intervention Framework illustrated in Figure 4.1 provides an updated outline for the cycles of improvement in this study following the decision to modify the original plan.

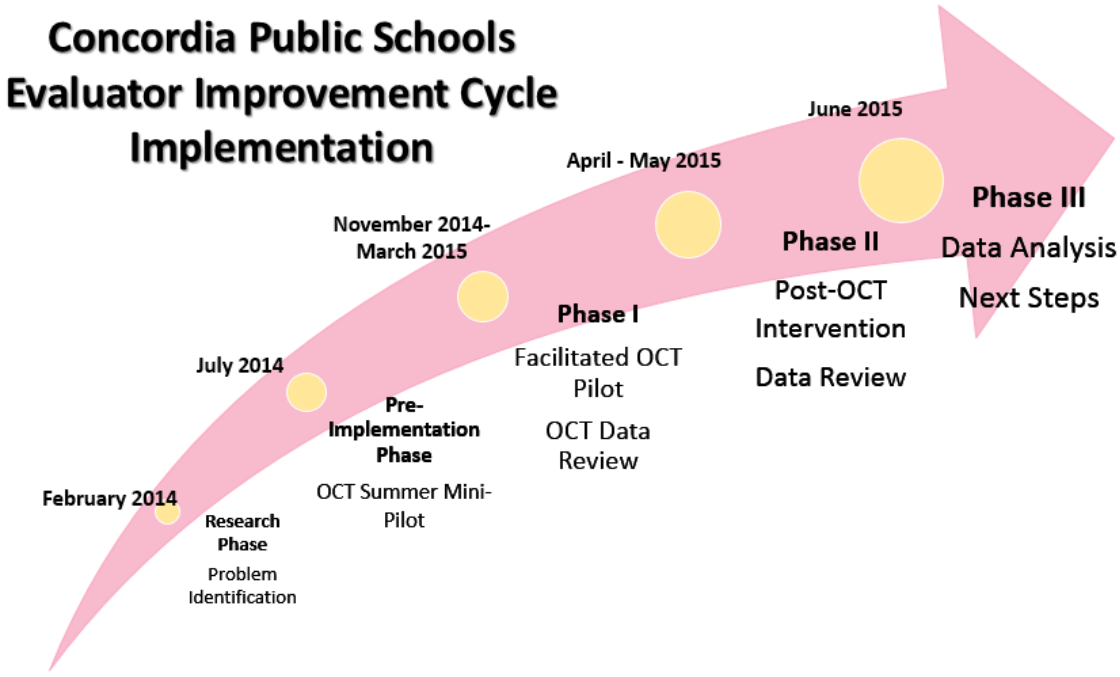


Figure 4.1. Modified project improvement cycle

Figure 4.1. Visual representation of the implementation timeline of the improvement project after revisions.

The intervention timeline spanned from February 2014 to June 2015. Final data from North Carolina's EVAAS became available in November 2015 for final review.

Implementation Plan Phases

This improvement plan was implemented in five clear phases – two pre-phases, which included research, data analysis, and the OCT Mini-Pilot, followed by two phases which were implemented during the 2014-15 school year and one phase which began in July 2015 and will continue throughout the 2015-16 school year. The outline in Figure 4.2 depicts the phases in each stage of the improvement project.

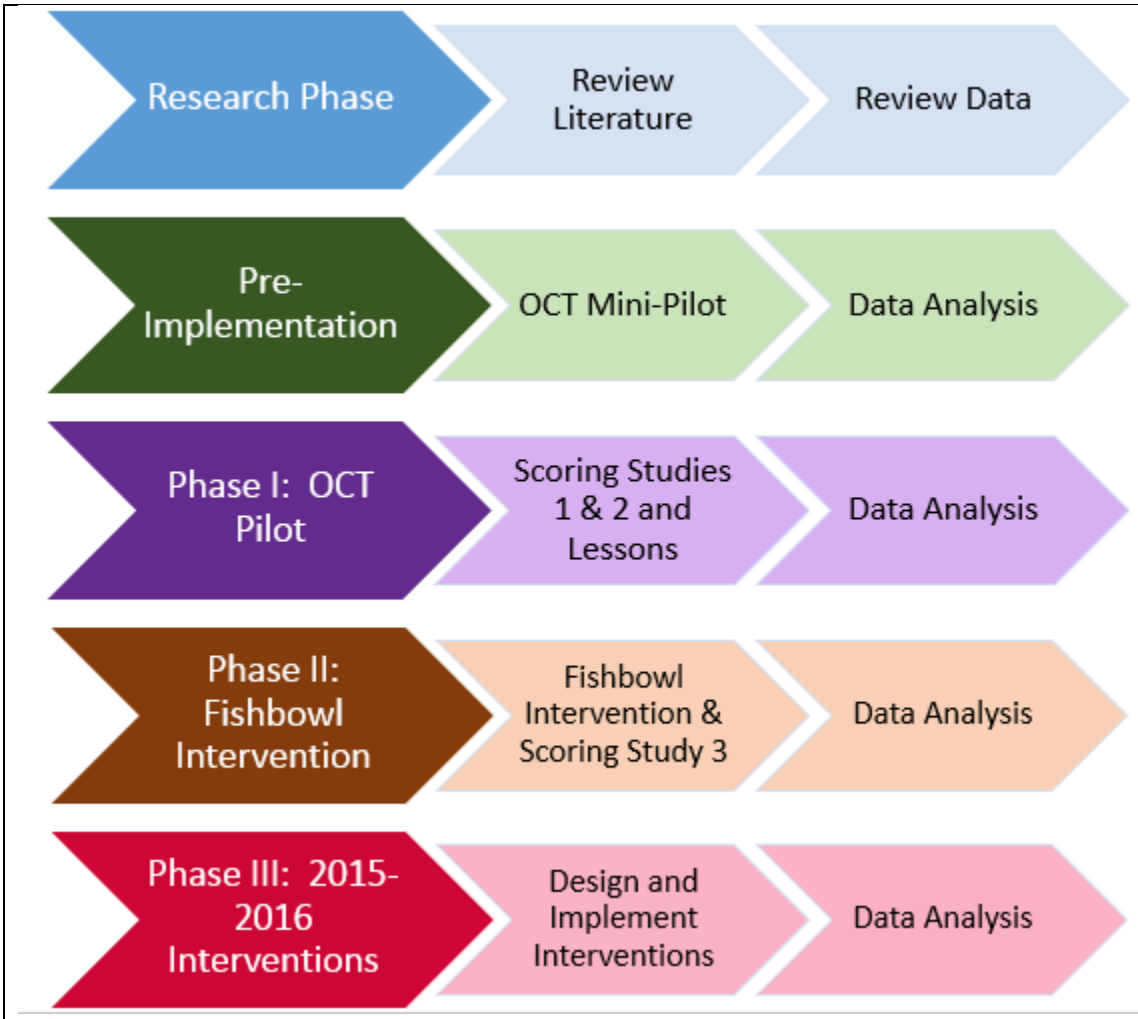


Figure 4.2. Improvement project phases of implementation

Figure 4.2. Breakdown of each of the five phases of the improvement project from research phase through post-project interventions to be implemented during 2015-2016.

Realignment of Goals for the Project

Phase I goals

- Concordia Public Schools evaluators and district leaders will participate in the 2014-15 OCT Pilot so that target agreement will increase by March 2015 as measured by a comparison of Scoring Study 1 and Scoring Study 2.
- Concordia Public Schools evaluators and district leaders will participate in the 2014-15 OCT Pilot so that scoring bias will decrease by March 2015 as measured by a comparison of Scoring Study 1 and Scoring Study 2.
- Concordia Public Schools evaluators and district leaders will participate in the 2014-15 OCT Pilot so that rater discrepancy will decrease by March 2015 as measured by a comparison of Scoring Study 1 and Scoring Study 2.

Phase II goal

- Concordia Public Schools evaluators and district leaders will participate in the Fishbowl Intervention so that target agreement will increase on the elements selected for the intervention by May 2015 as measured by a comparison of focus elements in Scoring Study 2 and Scoring Study 3.

Overarching project goal

- Concordia Public Schools principals will participate in the 2014-15 OCT Pilot and subsequent interventions so that the correlation between principal ratings on Standards 1-5 of the NCEES Rubric and Standard 6 will increase by October 2015 as measured by a comparison of EVAAS data from the 2013-14 school year and the 2014-15 school year.

Project Phases

The five phases of the project have been captured in the PDSA Cycles in Figure 4.3.

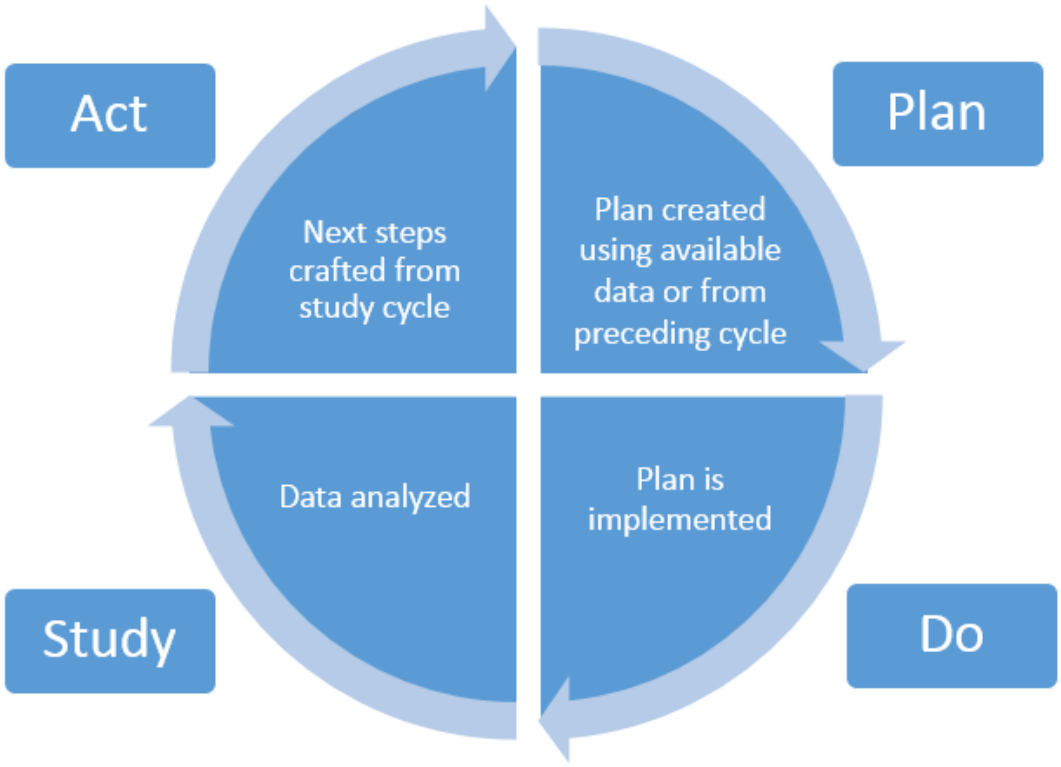


Figure 4.3. PDSA cycle outline

Figure 4.3. Overview of the PDSA cycle used to guide the improvement project phases.

Pre-Implementation Phase: Observation Calibration Training Mini-Pilot

Of all of the initiatives in Concordia Public Schools, the initiative that most impacted the interventions and assessments for this project was the OCT Mini-Pilot. After a review of the research and of the district's EVAAS correlation data, the district's

leadership team had investigated opportunities to improve the quality of evaluation. The OCT Mini-Pilot provided that opportunity.

Concordia Public Schools was selected to participate in the NCDPI OCT Mini-Pilot from June 30 – July 11, 2014. The NCDPI used Race to the Top (RttT) funds to develop an online platform to provide structured and focused observer training in an effort to improve rater accuracy and agreement. Because North Carolina has no evaluation certification system, the goal was originally to develop a platform that may become a component of a state-wide certification system for aspiring principals and evaluators. The discussions that followed the mini-pilot and concerns raised by the mini-pilot led to a PDSA cycle that resulted in the intervention design for this disquisition (Figure 4.4).

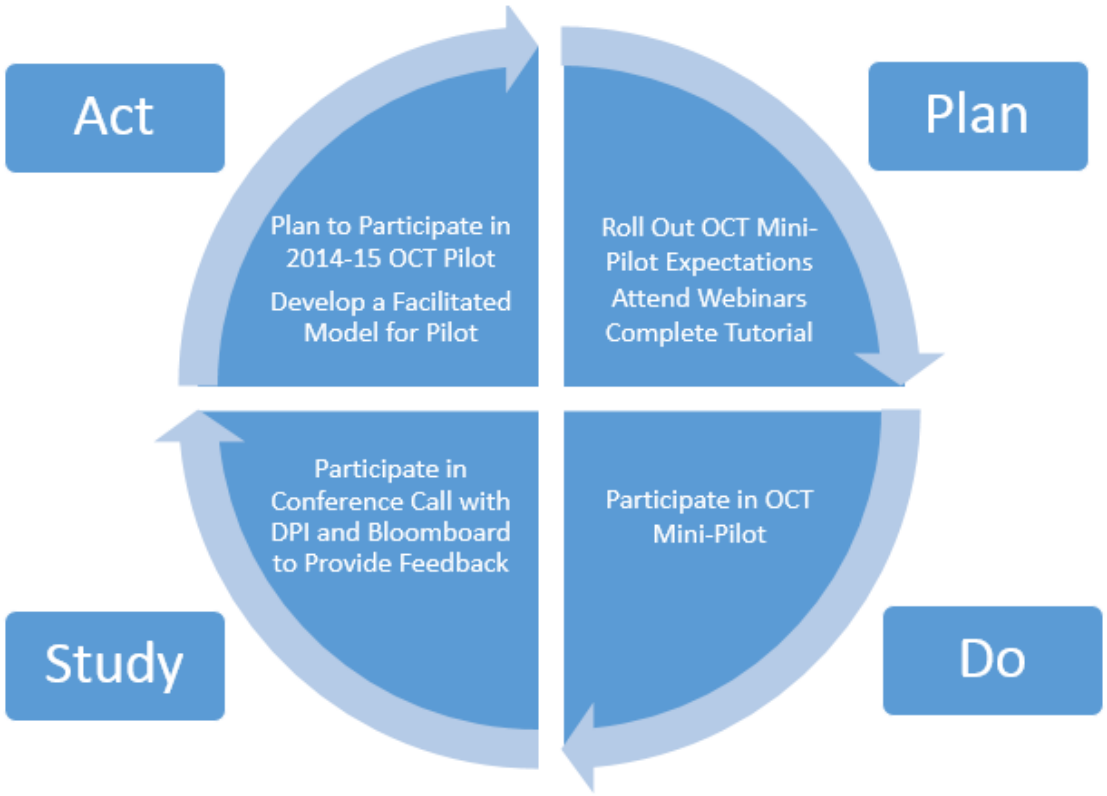


Figure 4.4. Pre-Implementation Phase PDSA Cycle – OCT Mini-Pilot

Figure 4.4. During the pre-implementation phase, seven Concordia evaluators participated in the NCDPI OCT Mini-Pilot. The results of this pilot informed Phase I.

To prepare to take part in the OCT Mini-Pilot, Concordia principals and curriculum directors attended a webinar co-facilitated by the NCEES Consultant at NCDPI and BloomBoard. Participants were provided with a rationale behind the development of the OCT Mini-Pilot as well as a tutorial on how to access and interact with Observation Engine, the online platform used to facilitate the mini-pilot. Observation Engine allows evaluators to access videos of real classrooms where they can observe instruction, rate teachers on specific standards and elements, and receive

immediate feedback on their performance. The participants were provided with a timeline to complete the mini-pilot. The mini-pilot consisted of one full-length classroom video that participants watched and rated all 17 observable elements of the North Carolina Professional Teaching Standards as well as five short (two to five minute) videos where participants rated one element of one of the five North Carolina Professional Teaching Standards. Participants were informed that they would be asked to provide feedback on their experience with the platform and the content of the professional development tool.

This pilot was designed for a small sample of evaluators and district leaders to review and provide feedback on both the content of the pilot and the functionality of the online platform where the contents of the pilot were housed, Observation Engine, provided by BloomBoard. Six of Concordia's seven principals participated in the OCT Mini-Pilot along with the CAO. According to the NCEES Consultant from NCDPI, participants from CPS and ACS, the two districts who participated in the OCT Mini-Pilot, engaged in two online discussions and provided feedback to the NCDPI as well as to the consultants from BloomBoard on July 11-12, 2014. Quantitative data from the OCT Mini-Pilot assessments as well as qualitative data gathered from the online discussions with NCDPI provided insight into participant scoring practices, perceptions of the OCT, and participants' self-efficacy as related to evaluation practices.

Components of the scoring study report. According to Empirical Education, "Observer agreement indicates the percent of an Observer's scores matching the target scores exactly." An observer is "discrepant" when he or she scores two or more ratings above or below the target. Furthermore, Observation Engine also completes a calculation

to determine whether or not a participant's responses demonstrate a statistically significant bias to score higher or lower when the participant disagreed with the target score. However, bias is reported only if there is a statistically significant (95%) chance that the participant's scoring pattern indicates a propensity to rate higher or lower.

Raw data from the one full-length Scoring Study (SS) of the pilot suggested that Concordia Public Schools evaluators could use further interventions to support target agreement and rater accuracy. Table 4.1 provides data collected from the SS during the OCT Mini-Pilot.

Table 4.1

Concordia OCT Mini-Pilot Scoring Study Results

Criteria Assessed	Concordia Public Schools Performance Results (n = 7)
Mean percent target agreement	44
Mean percent discrepant	12
Number of participants demonstrating scoring bias	3

Data in Table 4.1 was provided by Empirical Education and was analyzed using their proprietary software.

Data from the SS suggested that Concordia evaluators were not evaluating accurately. Furthermore, 12% of all ratings were discrepant, higher or lower than the target by two or more ratings. Forty-three percent of evaluators also expressed scoring bias on the SS. Two evaluators exhibited positive bias while the third expressed negative bias. These data indicated a need for further interventions to support evaluator growth.

Webinar participation and results. Following the pilot, participants from Concordia Public Schools and Appalachian County Schools took part in a webinar to

provide feedback to the NCDPI and BloomBoard regarding the content of the OCT Mini-Pilot and the Observation Engine platform developed by BloomBoard. As a participant in the pilot, I was able to provide feedback as well as record the feedback provided by other participants in both school districts that participated. Simmons (2014) developed a report on the feedback participants provided during the mini-pilot to director of Educator Effectiveness at the NCDPI. The report included the following feedback:

- Some participants disagreed with scores and wanted to know who had developed the master ratings.
- Participants asked for more feedback on what the teacher could have done differently to improve his/her rating and suggested that coaching recommendations for the teacher would be beneficial.
- Many participants expressed their surprise that they did not select the correct rating on many of the exercises.
- Participants expressed a need for clarity and understanding about master ratings and the process for developing those ratings.
- Several mentioned they would appreciate the ability to collaborate and discuss the videos as a group and process the videos and ratings together.
- Participants found that some video clips were too short to provide clear evidence of the standards.
- Participants liked the introductory information that helped them understand the content and context of the lesson.
- Participants saw the benefit of the tool for new or aspiring administrators and for specific support on individual areas for improvement.

- Participants provided confirmation of the need for the training provided by NCDPI through Observation Engine.

Feedback suggested that principals did like the immediate feedback from the pilot, the format, and the rationale. Most importantly, they stated that they wanted to experience training such as this in a collaborative setting. They admitted that they were surprised that they did not score more exercises correctly and that they saw the value and need for calibration training. These comments led to the implementation team meeting to design changes to the next cycle of support for evaluators. The OCT Mini-Pilot raised more questions than it provided answers in terms of the seven Concordia participants who took part in the study.

Personal experience. After participating in the mini-pilot and participating in the July 11 focus group webinar where other participants shared their views, I had the opportunity to reflect on my own experience as well as hear what principals thought and valued about the experience. Most importantly, I was able to hear more about what they wanted and needed out of evaluation support and professional development. This first-hand experience as a participant allowed me the unique opportunity to develop next steps for Concordia principals based on their feedback as well as my own involvement in the mini-pilot.

Phase I Intervention: Facilitated OCT Pilot

Following the July 2014 OCT Mini-Pilot, the NCDPI offered LEAs across the state the opportunity to take part in the OCT Pilot from November 2014 through June 2015. NCDPI stated that the purpose of the OCT Pilot was to collect data to determine whether or not target agreement improved and both scoring bias scoring discrepancy

decreased as a result of participation. The feedback from CPS principals who participated in the OCT Mini-Pilot indicated that they liked the opportunity to focus on one element at a time, that they were surprised they did not rate more lessons correctly, and that they felt a more collaborative model would be more effective. Participants agreed that completing the mini pilot individually was challenging. Mr. Thompson, principal at Central Concordia Elementary School, said that he wanted to ask others what they thought of the videos when he was confused about a rating. Having been a participant myself, I felt the same frustration upon reviewing a video, rating the video, and then discovering my rating was incorrect. As a matter of face, both OCT Mini-Pilot participants in Appalachian County Schools as well as those in CPS admitted to seeking out other participants to view videos together so that they could discuss what they saw. Mr. Barton, principal of Appalachian County High School, shared in the July 11 webinar that he and his assistant principal watched the videos together and then discussed them back and forth after viewing the first video and not selecting the target rating. As a participant in the OCT Mini-Pilot, I had the same experience. I was conflicted after viewing the first video and then asked for the opinions of my colleagues as well. I needed more clarity and feedback than the short rationale provided in the Bloomboard platform.

This feedback led the implementation team in Concordia Public Schools not only to request to be participants in the 2014-15 OCT Pilot but also to develop a partially-facilitated model and an implementation timeline. The implementation team worked jointly with the NCDPI NCEES Consultant to develop this model based on feedback from both Concordia's participants in the OCT Mini-Pilot and the principals from

Appalachian County Schools. The PDSA cycle for Phase I was based on the data, personal experiences, and outcomes from the Pre-Implementation Cycle. Figure 4.5 provides an overview of the PDSA cycle for Phase I.

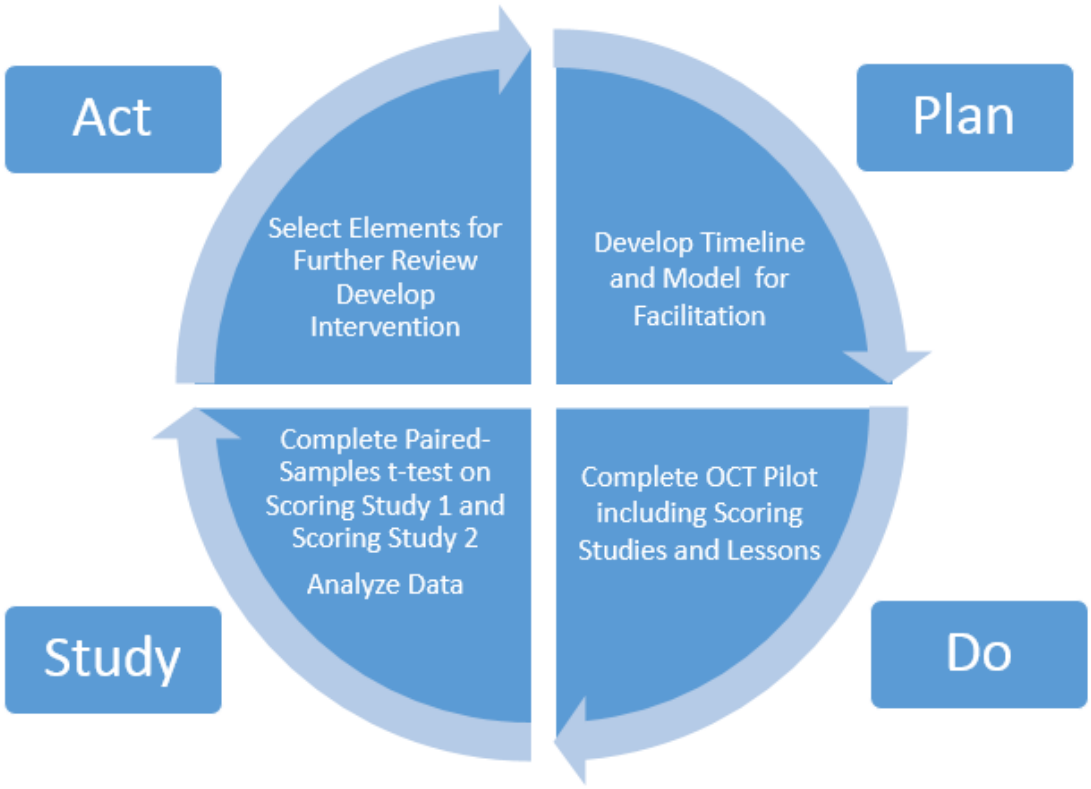


Figure 4.5: Phase I PDSA Cycle - OCT Pilot Implementation

Figure 4.5. The Phase I PDSA cycle implementation was devised based on data and feedback from the Pre-Implementation Phase (OCT Mini-Pilot). This iteration of the project lasted from November 2015 – March 2016.

The PDSA Cycle was designed to ensure that the plan for implementation was based on the data review and actions in the previous PDSA cycles which were informed by research and data analysis as well as the OCT Mini-Pilot participation. The cycle

provided a clear road map for the implementation and design teams to ensure that steps were developed as a direct result of data and participant needs, that appropriate data was collected, and that actions would be taken to inform the next PDSA cycle in Phase II.

Timeline and facilitation protocols. One important element the implementation team took under consideration was time. With two full-length scoring studies and 34 short videos, providing a clear timeline for implementation was crucial. Principals are busy with both operational and instructional responsibilities, and participation in the OCT Pilot did require a great deal of time. Principals admitted that they were concerned about the amount of time viewing the videos would take. Based on my own experience in the mini-pilot, I found myself feeling pressured to get the videos watched during the mini-pilot window. In order to alleviate stress and potentially get better participation, the implementation team asked for feedback from principals regarding support for developing a local timeline for OCT implementation. The implementation team then developed a viewing timeline to ensure that participants not only had a gauge on when to watch the videos provided in the OCT pilot but also had multiple opportunities to view the two full-length scoring studies and 12 of the 34 videos together and then engage in discussion about the lessons observed, master ratings, and specific wording of the element observed. The participants were provided with a timeline that spanned over four months. (See Table 4.2).

Table 4.2

Phase I - Concordia OCT Pilot Implementation Timeline

Date	Assignment/Event	Facilitation Method
November 18	OCT Kickoff and Overview of Timeline with NCDPI NCEES Consultant View SS1, rate, and discuss master ratings with NCEES Consultant	Facilitated
December 3	Review of timeline Lesson 1a	Facilitated
December 3 - January 7	Lesson 2a Lesson 2b	Completed Individually
January 7	Lesson 2c	Facilitated
January 8 – January 12	Lesson 2d Lesson 3a	Completed Individually
January 12	Module 3b	Facilitated
January 13 - February 4	Lesson 3c Lesson 3d	Completed Individually
February 4	Module 4a	Facilitated
February 5 - February 17	Lesson 4b Lesson 4c	Completed Individually
February 17	Module 4d	Facilitated
February 18 - March 4	Lesson 4e Lesson 4f	Completed Individually
March 4	Lesson 4g	Facilitated
March 5 - March 25	Lesson 4h	Completed Individually
March 25	Scoring Study #2	Facilitated

During September and October 2014, the implementation team met weekly to plan for implementation. The team developed a timeline with benchmarks for participants, created a YouTube video with instructions for logging into the Observation Engine platform on the BloomBoard website where all lessons were housed, and developed a protocol for lessons to be facilitated in a face-to-face setting. The implementation team also developed a protocol for facilitating the six lessons that would be viewed collaboratively during principals' meetings (Appendix I). The team also determined that the most logical time to facilitate face-to-face lessons would be during principals' meetings. Concordia held two principals' meetings per month during the 2014-2015 school year. Each principals' meeting was scheduled from 8:30 am – 12:00 pm. The first meeting of the month was devoted to operations, and the second meeting was slated for curriculum and instruction. The director of elementary education suggested that all principals, all curriculum directors, and the director of human resources take part in the OCT Pilot. These stakeholders attended all CPS principals' meetings and were directly responsible for curriculum and instructional support, including evaluation support. The director of high schools shared that he had not been a principal since North Carolina adopted the NCEES rubric and that participation in the pilot would help him support principals as well. Principals' meetings provided an appropriate, uninterrupted time to engage in discourse around videos and collect informal qualitative data.

The implementation team worked with all of Concordia's executive leadership to develop a plan to schedule 30-40 minutes of six principals' meetings for observing one lesson together. Each lesson consisted of two videos, and meeting time would be used to review the element to be rated, view and rate the teacher of each lesson individually, and

then discuss the ratings as a group. The timeline in Table 4.2 was developed so that principals would view one lesson together and two lessons on their own between principals' meetings. One third of the lessons were facilitated during principals' meetings while two thirds of the lessons were completed by participants during their own time between principals' meetings.

Kickoff. The implementation team decided that a face-to-face formal kickoff of the pilot would be important so that participants had the opportunity to ask questions and provide feedback on the timeline and process. The NCDPI NCEES Consultant worked closely with the implementation team and offered to attend the November 18, 2014 OCT Pilot kickoff meeting in Concordia to meet with principals and share information regarding the pilot as well as facilitate a discussion following SS1. During the meeting, the NCDPI NCEES consultant facilitated an overview of the pilot, administered the first Scoring Study, and assisted participants as they logged into the OCT Pilot platform. After all participants logged into the BloomBoard platform, they were provided with a paper copy of the North Carolina Educator Evaluation Manual. The group viewed the video collectively, and then without discussion, each participant rated the seventeen elements in the Bloomboard platform. Participants were encouraged to use the paper copy of the rubric to mark their ratings to reference during the discussion following the exercise. Immediately following the video, participants were asked to enter their ratings into SS1 in Observation Engine and rate all 17 observable elements as one of the following:

(0) – Not Observed

(1) – Developing

(2) – Proficient

(3) – Accomplished

(4) – Distinguished

Critical role of post-observation facilitated discussion. Although participants did receive immediate feedback on the lessons portion of the OCT design, no feedback was provided on the Scoring Studies. However, the NCEES Consultant had access to the master ratings for the 17 elements, and she facilitated a discussion with participants regarding the master ratings, language of the standards, and misconceptions observers had. During the impromptu discussion following the review of elements, some participants shared concerns about the master ratings and their strong convictions about the wording of the standards. The remarks echoed some of the same concerns participants shared during the OCT Mini-Pilot. One principal shared that she was not convinced that the ratings were accurate and questioned who the master raters were. The NCEES Consultant shared information regarding the validation process and the quality of the ratings by a team of master raters.

Having a respected NCDPI leader provide this level of customized support and feedback gave principals an opportunity to share their observations, thoughts, and concerns about ratings assigned by master raters on the pre-assessment and get immediate feedback to their questions. All agreed that the process was valuable and that the discussions around the wording of the standards was helpful to their practice. One principal shared that this type of activity was among the best the district had provided in terms of support with evaluation.

Structure, format, and implementation of OCT lessons. Over the course of four months, Concordia participants reviewed 34 short (2-8 minutes) videos within 17 lessons that were each designed to display one of the 17 observable elements of the North Carolina Professional Teaching Standards. A timeline for lesson review was provided to chunk the lessons in small increments that participants could manage over the course of the pilot. Of the 17 lessons, participants viewed and rated 11 lessons individually and viewed six lessons as a group. During the group sessions, the facilitator began discussions by asking if there were questions, concerns, or comments regarding the lessons that participants had completed on their own since the last face-to-face session. Often, participants shared confusion, disagreement with ratings, or questions they had after viewing a particular video. Short, impromptu discussions took place where participants exchanged ideas about the lessons, video quality, supplementary materials, or feedback provided within Observation Engine.

Prior to completing lessons collaboratively during principals' meetings, participants received a hand out with the element that participants rated during the lesson (See Appendix J). The facilitator, either the CAO or director of elementary education, used the protocol for facilitation that was developed by the design team to facilitate the discussion. The participants logged into Bloomboard. The facilitator handed out the element to be observed to each participant. The participants then silently read the element and each descriptor and engaged in a micro-discussion regarding "look-fors" and wording of the standard. Participants were directed to review all supplementary materials provided in the lesson. Supplementary information included the grade level, content area, and context of the lesson. Supplementary documents often included lesson plans,

student work, handouts, and other documents pertinent to gaining a greater understanding of the context of the videos. Participants could open these documents to glean more information about the lesson and the teacher's objective.

After each participant had the opportunity to review materials and documents, the group viewed the short videos together without discussion. After viewing, each participant was asked to rate the teacher individually in Bloomboard and submit ratings electronically into Observation Engine. Participants then immediately received both the master ratings and feedback through the Bloomboard Observation Engine platform. One of the major differences between SS1 and the 34 short videos was the coaching nature of these exercises. Unlike the scoring studies, participants received immediate feedback regarding their ratings. The master rating was provided with a brief explanation of how the particular classroom lesson should have been rated and why. For example, if a teacher was rated "proficient," feedback might include what the teacher said or did or specific materials the teacher used with students. The nature of the questioning or other pedagogical strategies may be explicitly explained to defend the rating.

After all participants had rated the teacher in the video lesson and received feedback, the facilitator led a discussion around the descriptors, misconceptions, the language and wording of the specific element under review, instructional practices observed, limitations of the videos and materials, evaluation practice, and insights gleaned from the exercise. Each of the two discussions lasted from five to ten minutes. The same format was followed for all six lessons that participants completed collaboratively in the facilitated model.

Collaborative participation in Scoring Study 2. The implementation team developed a timeline that was provided to all participants from Concordia that indicated that all lessons should be completed before March 25. On March 25 during the Curriculum Principals' Meeting, participants who were present followed the same format for lesson viewing to observe SS2. The facilitator provided a copy of all 17 observable elements on the NCEES Rubric and guided participants to log into Observation Engine to input their ratings. Only seven of the 12 Concordia participants in the OCT Pilot attended this meeting due to a variety of reasons. The seven participants who were present viewed the full-length video in the platform together, rated the 17 observable elements separately, and then discussed the video. Participants were frustrated that master ratings were not provided to the participants upon submission of ratings. The facilitator reminded the participants that only because the NCDPI NCEES Consultant was with us for SS1 did we have access to the master ratings. Participants were anxious to know how they performed and whether or not they improved. Data was unavailable at the time, but the CAO assured the participants that once the data was available, all participants would have a chance to review the findings. The five participants who were not in attendance for the March 25 SS2 viewing and rating session were asked to complete SS2 prior to April 8. All five completed SS2 between April 1 and April 8, 2015 on their own time.

In May 2015, Bloomboard did provide an analysis for Concordia Public Schools comparing the results of SS1 and SS2. However, observers were not identified in the analysis. After receiving these results, the design team met to determine possible next steps to facilitate further improvement based on the holistic raw data.

Analysis of Phase I results. The design team developed four goals to drive the work behind Phase I of the intervention. Three of the goals were related to findings from the results of SS1 and SS2 in the OCT Pilot. SS1 provided a benchmark assessment of participant target agreement, target discrepancy, and scoring bias in November 2014. SS2, administered in March 2015, provided the opportunity for both the implementation and design teams to analyze the impact that participation in the OCT Pilot had on observers as individuals and collectively. All three measures were analyzed and evaluated through a paired samples t-test. Table 4.3 provides the analysis of that data.

Table 4.3

Scoring Study 1 and Scoring Study 2 Comparison for Concordia Public Schools

Criteria Assessed	Scoring Study 1 (n = 12)	Scoring Study 2 (n = 12)
Mean percent target agreement	49.6	54.9
Mean percent target discrepant	8	3*
Number of participants demonstrating scoring bias	7	2*

Note: *Difference from SS1 results statically significant at the $p<.05$
Data in Table 4.3 was provided by Empirical Education and was analyzed using their proprietary software.

Goal 1 results. Goal 1: Concordia Public Schools evaluators and district leaders will participate in the 2014-15 OCT Pilot so that target agreement will increase by March 2015 as measured by a comparison of Scoring Study 1 and Scoring Study 2**.** Based on the results from a paired samples t-test, target agreement did increase among the Concordia OCT Pilot participants from 49.6% to 54.9% as measured by a comparison of SS1 and SS2. In terms of target agreement in SS1, out of 17 possible responses, participants averaged having 8.58 incorrect responses. However, in SS2, participants

averaged only 7.66 incorrect responses. Nine of the 12 Concordia participants who participated in the pilot completed all 34 lessons including both scoring studies. A deeper analysis of the raw data from the study indicated that six of the 12 participants demonstrated improvement in agreement with target scores from SS1 to SS2. Four participants demonstrated a decline in target agreement. Two of the four participants who demonstrated a decline did not complete all of the 34 lessons within the OCT Pilot intervention.

Over the course of the pilot, participants shared their insights freely during facilitated discussions. On March 25, the CAO asked OCT Pilot participants to share their general impressions of the pilot. However, five of the seven participants were not present for this meeting. Mr. Hines, the principal of Concordia High School, stated that evaluation is not only about what you see during a lesson but also about what you know has happened prior to the observation. He added that the fact that there is no prior relationship or background in the videos makes it more difficult to evaluate. However, he did add that this fact does force the participants to find evidence.

Two elementary principals, Mrs. Felton and Mr. Thompson, discussed whether or not it would be advantageous for North Carolina to use the OCT as a component of principal licensure programs or as a component of the first year of an assistant principalship or first year of the principalship as a certification tool. Their concerns were that they would not be able to pass the OCT because it had proven to be so challenging. Mr. Thompson also said that he loved watching the videos as a group and discussing them, but he did not like doing them at home on his own. Both the Director of Elementary Education and Mr. Hines shared the same view. Both agreed that they found

value in the group discussions. Mr. Thompson added that after he watched a video and received feedback, he wanted to turn to his colleagues and talk about the ratings and the feedback.

Goal 2 results. Goal 2: Concordia Public Schools evaluators and district leaders will participate in a partially facilitated model of the 2014-15 OCT Pilot so that scoring bias will decrease by March 2015 as measured by a comparison of Scoring Study 1 and Scoring Study 2. Data from SS1 indicated that 58.3% of participants' scoring trends showed statistically significant scoring bias. However, data collected from SS2 indicated that only 16% of Concordia participants demonstrated scoring bias after completing the 17 target lessons. This finding suggests that the partially-facilitated model improved rater accuracy and reduced scoring bias. A paired samples t-test concluded that there was a statistically significant improvement at the $p<.05$ in scoring bias.

Goal 3 results. Goal 3: Concordia Public Schools evaluators and district leaders will participate in the partially facilitated model of the 2014-15 OCT Pilot so that rater discrepancy will decrease by March 2015 as measured by a comparison of Scoring Study 1 and Scoring Study 2. Six of the 12 participants demonstrated improvement in target discrepancy from SS1 to SS2 as well. During SS1, participants ranged from 0% discrepant to 24% discrepant with a mean rating discrepancy of 8%. On SS2, participants ranged from 0% discrepant to 12% discrepant with a mean rating discrepancy of 3%. Again, a paired samples t-test concluded that there was a statistically significant improvement at the $p<.05$ in rater discrepancy. The improvement is particularly significant due to the small sample size of the group.

In a comparison of all three scoring studies, Empirical Education worked with Concordia Public Schools to analyze the data in a research brief. Only Element 2b provided difficulty for the participants and was the only element that participants demonstrated large deviations from the target score. Because discrepant scores were spread across all other elements, the data suggest that completing the OCT Pilot did support participants in their development of an essential understanding of the NCEES Rubric (Empirical Education, 2015b).

Conclusions: Collaborative versus independent learning. The comparative data from SS1 and SS2 suggests that the facilitated model developed and implemented by Concordia's OCT Pilot Implementation Team was effective. Further, participants seem to have developed a deeper understanding of the North Carolina Educator Evaluation Standards and the seventeen observable elements within the tool. The data is particularly significant due to the small sample size in the study. Group performance improved on each of the three measures that the studies assessed. Target agreement improved from 49.6% to 54.9%. Percentage of discrepancy declined 5% overall, and the number of participants who exhibited scoring bias decreased substantially.

One interesting aspect of the project that we did not consider was the effect attendance played on the participants. When Empirical Education provided the first batch of data based on the participants' performance on SS2, five participants had not completed the scoring study because they were absent from the group viewing session on March 25, 2015. After the five remaining participants completed the scoring study independently between April 1 and April 8, 2015, we asked Empirical Education to reanalyze the data and provide us with a new report. We noticed that the target

agreement dropped from those two reports. This realization prompted me to solicit the raw data from Empirical Education and run a paired samples T-test to determine whether attendance made a statistically-significant impact on target agreement and scoring bias. In Table 4.4, a comparison between the seven participants who completed SS2 as a group in a controlled setting and the five who completed the study on their own suggests that the environment may be a factor in participant performance.

Table 4.4

Comparison of Scoring Study 2 Viewing Collaboratively versus Viewing Alone

Criteria Assessed	Scoring Study 2	
	Viewing Collaboratively (n = 7)	Working Alone (n = 5)
Mean percent target agreement	59.6	48.4
Average targets missed	6.8	8.8
Mean target discrepant	2.6	3.6
Number of participants demonstrating scoring bias	1	1
Mean percent difference between SS1 & SS2	7.7	-2.6

Data in Table 4.4 were provided by Empirical Education and was analyzed using their proprietary software.

A deeper review of the data reveals that of the five participants that were not present to complete SS2 as a group, their observer agreement with target scores dropped from SS1 to SS2 by an average of 2.6%. Participants who were present to view SS2 together increased their observer agreement with target scores from SS1 to SS2 by an average of 7.7% and answered correctly on an average of two more items than those who worked alone. This prompted the implementation team to conduct a further review of variables related to collaborative vs. independent viewing. Further analysis indicated that

six of the 12 participants were present to view both SS1 and SS2 together. Five participants were not present for either principals' meeting where SS1 or SS2 were viewed. All five completed the viewing and scoring studies on their own time independently. One participant was not present for the collaborative viewing of SS1 but was present for SS2. Tables 4.5a, 4.5b, and 4.6 provide the comparison of the performance on SS1 and SS2 of the participants who were present for both assessment sessions and those who were not present for either of the scoring studies.

Table 4.5a Comparison of Target Agreement Performance of Participants Who Completed both Scoring Studies in a Controlled Environment

	SS1		SS2						
	M	SD	M	SD	Difference	t(5)	p	95%CI	Cohen's d
Agreement	44.00	9.85	56.83	11.46	12.83	-1.920	.113	[-30.01, 4.35]	2.60 {The formula for Cohen's d is $M_1-M_2/\sqrt{((s_1^2+s_2^2)/2)}$}

Six participants completed both SS1 and SS2 in a controlled environment. I measured their target proficiency on SS2 and SS3. On SS2, participant performance was below 50% accurate (M=44.00, SD=9.85). After the intervention, the six who viewed and rated SS3 in a controlled environment improved their target accuracy (M=56.83, SD=11.46) with a mean difference of 12.83. Further analysis with a paired-samples T-test revealed the difference between performance on these selected elements from SS1 to SS2 was not statistically significant, t(5)=-1.920, p=.113. However, this group did improve their performance by over 12%, whereas the group that did not complete either training in a controlled environment demonstrated statistically significant decline in target agreement by -7.2%. These findings are important to consider when developing interventions to support improvements in evaluation practices. It is also important to note that target

discrepancy decreased from 8% to 3% among these six participants and that scoring bias decreased from 3 participants (50%) to 1 participant (16.6%). This decline was statistically significant at $p<.025$. The scoring bias results for the six participants who completed both scoring studies in a controlled environment can be found in Table 4.5b.

Table 4.5b Scoring Bias of Participants Who Completed both Scoring Studies in a Controlled Environment

	SS1		SS2						
	M	SD	M	SD	Difference	t(5)	p	95%CI	Cohen's d
Scoring Bias	-.50	.54	.16	.40	0.66	-3.162	.025	[-1.20, -1.24]	2.60 {The formula for Cohen's d is $M_1-M_2/\sqrt{((s_1^2+s_2^2)/2)}$}

Five of the Concordia participants completed both scoring studies in an independent environment. Table 4.6 provides the comparison of target agreement performance from SS1 and SS2 for participants who were not present for either collaborative assessment session.

Table 4.6 Comparison of Target Agreement Performance of Participants Who Completed both Scoring Studies in an Independent Environment

	SS1		SS2						
	M	SD	M	SD	Difference	t(4)	p	95%CI	Cohen's d
Agreement	55.60	18.35	48.40	19.23	-7.2	3.207	.033	[-.96, 13.43]	2.60 {The formula for Cohen's d is $M_1-M_2/\sqrt{((s_1^2+s_2^2)/2)}$}

The data represented in Table 4.6 are a result of an unexpected finding. It appears that the target accuracy of participants who did not complete either scoring study in a controlled environment declined from SS1 (M=55.6, SD=18.35) to SS2 (M=48.40, SD=19.23) with a mean difference of -7.2. Further analysis revealed that this decline is

statistically significant (*p=.033)*. However, target discrepancy did decline from 8.4% to 3.6% for these participants, and scoring bias decreased from 60% to 20%. These outcomes are similar to the participants who were present for the collaborative viewing sessions. Such findings, specifically as related to target agreement, add additional support to the revelation that a structured environment and providing allocated time for evaluation support may have a positive impact on participants' accuracy.

Although the entire group of 12 participants did show improvement on all three measures in the study, the data in Tables 4.5a, 4.5b, and 4.6 suggest that one element of the facilitated model that impacts performance is the controlled environment. The protocol for viewing OCT Pilot videos during principals' meetings includes an emphasis on removing stimuli that would cause participants to lose focus. Norms include putting away computers and phones and focusing on only the video and the rubric provided by the facilitator. One variable that emerged during the study was simply the effect of the controlled environment on the participants.

While an inferential statistical test does not indicate statistical significance in the results of the group who completed both scoring studies in the controlled environment, the descriptive approach demonstrates a trend toward improvement in the group that met together to review SS2 as opposed to the participants who completed SS2 on their own. The small sample size of the two groups is one of the major reasons that statistical significance could not be established using a paired-samples t-test. It is important to note, however, that participants who completed SS2 on their own demonstrated a pronounced trend toward decline. Nine of the 12 participants completed SS1 in a controlled environment, and only seven of the 12 participants completed SS2 together on

March 25, while five of the 12 participants completed SS2 on their own between April 1, 2015 and April 8, 2015. This data was useful to the implementation team in designing further interventions. All raw agreement, discrepancy, and bias data from the OCT Pilot scoring studies can be found in Appendix K.

Although not a variable considered in the research proposal, the data uncovered in the Empirical Education raw performance data reveals a trend that the implementation team found significant for use in later iterations of the professional development intervention for principals. The group members who participated in both meetings where the group viewed the Scoring Studies together and then had time to rate separately improved their target agreement performance 13% from SS1 to SS2. The group's target discrepancy also improved, and only one participant exhibited scoring bias, as opposed to three during SS1.

The six participants who completed the both scoring studies together showed improvements on all three measures. However, the five participants who completed both scoring studies independently showed a decline in percent target agreement, although the other two measures did show improvement.

Phase II Intervention: Fishbowl

To develop a logical Phase II iteration of the intervention, the design team took into consideration both participant feedback and research regarding professional development as well as time constraints. Over the course of the OCT Pilot, participants shared their insights freely during facilitated discussions. On March 25, the CAO asked OCT Pilot participants to share their general impressions of the pilot in a group setting. Five of the seven participants were not present for this meeting. Mr. Hines, the principal

of Concordia High School, stated that evaluation is not only about what you see during a lesson but also about what you know has happened prior to the observation. He added that the fact that there is no prior relationship or background in the videos makes it more difficult to evaluate. Mr. Hines did add that this fact does force the participants to locate specific evidence to justify their ratings. Mrs. Pennington, principal at Concordia Academy, shared that the quality of the video made it challenging to really observe what the students were doing or what types of conversations they were having. All participants agreed that observing teachers in their classrooms would provide a better opportunity for more robust discussions and accurate rating due to the fact that observers would have the ability to move around the room, access documents and plans, and ask clarifying questions. These comments were taken into consideration when developing the Phase II intervention.

Consideration of adult learning theory. The design team, along with the implementation team, wanted to ensure that the Phase II intervention was purposeful and that it provided the best possible conditions for participants. To achieve this goal, the team reviewed the tenets of adult learning theory. According to Merriam (2001), Adult Learning Theory or Andragogy, developed by Malcolm Knowles, operates under five assumptions that describe the "adult learner as someone who:

- Has an independent self-concept and who can direct his or her own learning
- Has accumulated a reservoir of life experiences that is a rich resource for learning
- Has learning needs closely related to changing social roles
- Is problem-centered and interested in immediate application of knowledge
- Is motivated to learn by internal rather than external factors" (p. 5)

Adult Learning Theory instructs that professional developers "should involve learners in as many aspects of their education as possible and in the creation of a climate in which they can most fruitfully learn" (Merriam, 2001, p. 7). Andragogy hones in on the notion of ensuring any type of adult learning provides participants the opportunity to direct their own learning and that opportunities be learner-centered. The principles of andragogy include the fact that adults have a need to be involved in the planning and evaluation of their instruction. Adults believe that experience should provide the basis for learning activities. Learning is most interesting to adults when they see that the experience has immediate relevance to their work or personal life. Adults learn best when the experience is problem-centered rather than content-oriented.

When considering any professional development intended to change educator practice and impact student achievement, district leaders would be remiss without taking time to consider the significant role that Adult Learning Theory plays in the development plan. Throughout each stage of the professional development process, leaders must ensure that stakeholders understand the need for the professional development and have some input into the content, duration, time, and location of the professional development. The professional development itself must provide adults with multiple opportunities to share their experiences and expertise and apply their knowledge to the given task.

The design team discussed the tenets of high quality professional development, adult learning theory, and data from the OCT Pilot to design a Fishbowl Intervention, loosely based on the Socratic Method. Some members of the design team had previously studied Guskey's Five Critical Levels of Professional Development Evaluation and understood that high quality professional development must be measured by more data

than whether participants enjoyed the professional development (See Appendix L). The CAO's previous position was at the NCDPI as a professional development consultant. In that role, she studied Adult Learning Theory and tenets of Andragogy. Likewise, the NCEES Consultant had the same training as well as served previously as a national-level professional development facilitator. Because the nature of understanding the standards is based around wording, observation, and "look-fors," it was important that the Phase II iteration be designed around collaboration, an opportunity for all stakeholders to share thoughts, and knowledgeable facilitators who were considered experts. The CAO provided the design team with information about Socratic Seminars from previous trainings she had completed. A modification of the Socratic Fishbowl Method was selected by the design team as the vehicle for the next iteration of the intervention.

The Phase II intervention occurred on April 28, 2015 and consisted of Concordia principals and district leaders accompanying two experts from the Educator Effectiveness Division of the NCDPI into classrooms in small groups to conduct abbreviated observations. Evaluation of this phase occurred on May 19, 2015 through the administration of an additional scoring study (See Table 4.7).

Table 4.7

Phase II - Concordia Fishbowl Intervention Implementation Timeline

April 28	Review the Data from Scoring Studies 1 & 2 Fishbowl Intervention on Elements 1a, 2d, 4a, and 4d with NCDPI NCEES Consultant	Facilitated
May 19	Scoring Study #3	Facilitated

Two groups of participants simultaneously observed two classrooms. Participants, including facilitators, were asked to rate teachers on specific observable elements. After the observations, participants returned and participated in a Socratic-style "fishbowl" discussion led by the field expert from NCDPI. Participants discussed ratings, "look-fors," specific classroom examples, and questions they had about the element, language of the tool, or the observation process in general. Meanwhile, the remaining principals and curriculum directors sat in an outer circle and took notes on what they heard, questions they had, insights they gleaned from the discussion, and other important moments of confusion or clarity on the "Capture Your Thoughts" tool provided by the design team (Appendix M). After the discussion, the facilitator asked the participants in the outer circle to share their thoughts, moments of clarity, or questions with the group. Following these comments and ensuing discussion, the groups switched positions, and the group that was participating in the discussion became the note takers while the note takers in the outer circle became participants in the discussion. All participant notes were collected and coded using the In Vivo and Pattern coding methods. The session concluded with an opportunity for participants to reflect on what they gleaned from the exercise and what they needed next.

The PDSA Cycle in Figure 4.6 provides the key components of the Phase II iteration of the implementation plan.

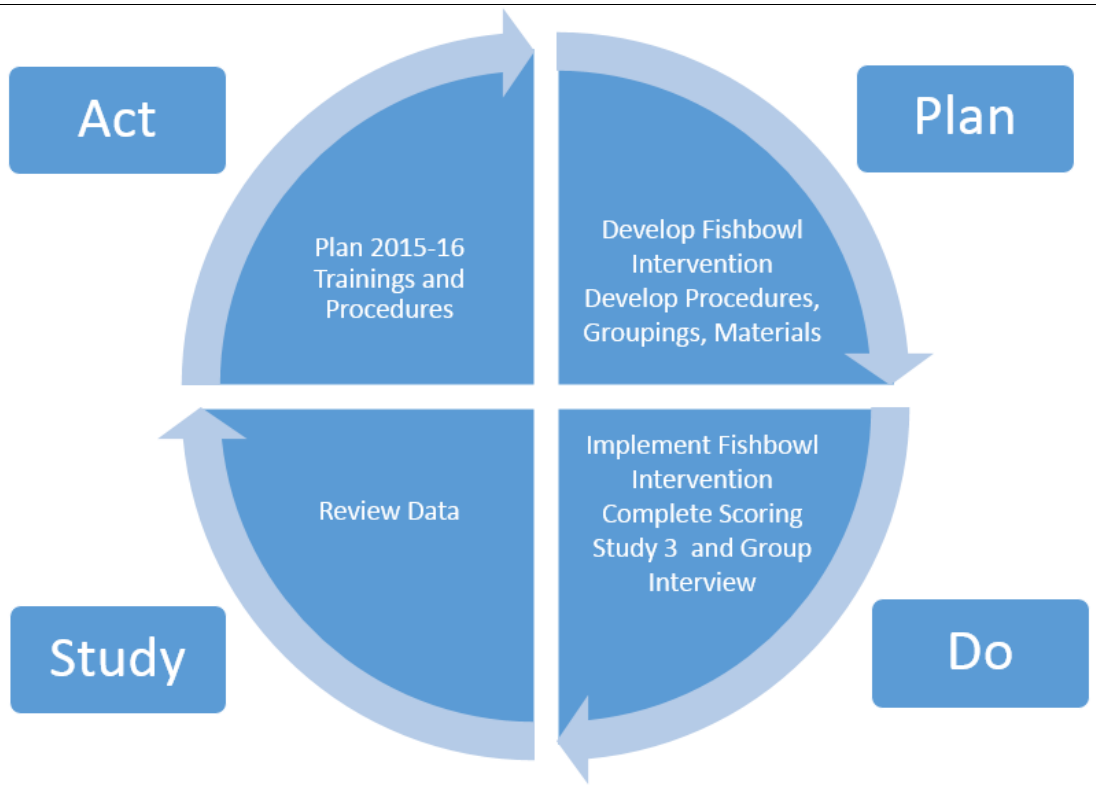


Figure 4.6. Phase II PDSA Cycle – Fishbowl Intervention Implementation

Figure 4.6. The Phase II PDSA cycle implementation was devised based on data and feedback from the Pre-Implementation Phase (OCT Mini-Pilot) and Phase I (OCT Pilot). This iteration of the project lasted from April 28, 2015 – May 19, 2016.

Group selection. At the time of group selection, the implementation team had access only to raw group data and no individual performance data. Group selections for the fishbowl were based solely on personality and job description. The team decided to ensure that principals and district leaders with primarily elementary experience were equally intermingled with participants with primarily secondary experience in order to provide multiple perspectives on the elements under review. Furthermore, personalities were taken into account. In an attempt to provide more opportunities for participation,

participants who traditionally shared more of their opinions and ideas in meetings and had presented themselves as more extroverted during discussions were equally distributed among groups.

Selection of expert facilitators. The NCDPI NCEES Consultant was a member of the design team, and she agreed to lead one of the Fishbowl discussions. She also volunteered another member of the NCDPI NCEES Team to lead the other group. Both of the facilitators had experience as successful principals. Between the two of them, they had experience at the elementary, middle, and high school levels as teachers and as administrators. The facilitators were selected in part due to their roles at NCDPI. Furthermore, both facilitators are considered NCEES experts and travel the state of North Carolina working with principals and district leaders to gain a greater understanding of the standards, the evaluation tool, and the evaluation process.

Selection of standards. The four elements the design team selected for the Fishbowl Intervention included Standards 1a, 2d, 4a, and 4d. These four elements were all targeted for specific reasons. After reviewing data from the mini-pilot, SS1, the 17 lessons, and SS2, the design team wanted to select at least one element in which the Concordia OCT Pilot participants had demonstrated growth, decline, and virtually no movement. An analysis of each element was conducted by reviewing the data available and selecting elements for review. A thorough analysis with anecdotal notes can be found in Appendix N. Because the Fishbowl observation would last only 20 minutes, the design team agreed that any more than four elements would be too many to evaluate appropriately. In the NCEES, Standards 1-4 contain elements that can be observed in the

classroom during observations. Table 4.8 depicts the elements that were selected for the Fishbowl Intervention.

Table 4.8

Comparison of Quantitative Measures from June 2014 – March 2015

Selected Standards	Mini-Pilot % Target Agreement (*n* = 7)	SS1 % Target Agreement (*n* = 12)	OCT Lessons % Target Agreement	SS2 % Target Agreement (*n* = 12)
Standard 1a: Teachers lead in their classrooms.	29	58	50 (*n* = 12)	42
Standard 2d: Teachers adapt their teaching for the benefit of students with special needs.	29	25	17 (*n* = 12)	33
Standard 4a: Teachers know the ways in which learning takes place, and they know the appropriate levels of intellectual, physical, social, and emotional development of students.	0	58	29 (*n* = 12)	17
Standard 4d: Teachers integrate and utilize technology in their instruction.	57	50	56 (*n* = 9)	67
Average Target Agreement	28.75	47.75	38	39.75

Standard 1a. Standard 1a is the only observable element of the five elements in Standard 1. The remaining elements are evaluated over time based on what a principal knows about a teacher's performance and what the principal and teacher document over the course of the school year. Standard 1 addresses teacher leadership. Often, teacher

leadership is demonstrated in a teacher's work outside the classroom with his/her colleagues, in professional learning communities (PLCs), facilitating professional development, involvement with school or district leadership teams, development of policies, using data to inform instructional decisions, etc. These practices are not observed in classroom instruction. Standard 1a provides an even more interesting challenge to the observer because of its 11 descriptors, only three are observable in the classroom. On the Rubric for Evaluating North Carolina Teachers, a checkmark to the left of an element indicates whether or not an element is observable. Figure 4.7 provides the visual representation of Standard 1a that evaluators see. Unlike all other observable elements in the NCEES, Standard 1a has no descriptor for the "Developing" rating. A classroom observer would only be able to observe the ratings "Not Demonstrated," "Proficient," "Accomplished," or "Distinguished."

Observation	Element Ia. Teachers lead in their classrooms. Teachers demonstrate leadership by taking responsibility for the progress of all students to ensure that they graduate from high school, are globally competitive for work and postsecondary education, and are prepared for life in the 21st century. Teachers communicate this vision to their students. Using a variety of data sources, they organize, plan, and set goals that meet the needs of the individual student and the class. Teachers use various types of assessment data during the school year to evaluate student progress and to make adjustments to the teaching and learning process. They establish a safe, orderly environment, and create a culture that empowers students to collaborate and become lifelong learners.				
	Developing	**Proficient**	**Accomplished**	**Distinguished**	**Not Demonstrated (Comment Required)**
		… and	… and	… and	
	❑ Understands how they contribute to students graduating from high school.	❑ Takes responsibility for the progress of students to ensure that they graduate from high school.	❑ Communicates to students the vision of being prepared for life in the 21st century.	❑ Encourages students to take responsibility for their own learning.	
	❑ Uses data to understand the skills and abilities of students.	❑ Provides evidence of data-driven instruction throughout all classroom activities.	❑ Evaluates student progress using a variety of assessment data.	❑ Uses classroom assessment data to inform program planning.	
✓		❑ Establishes a safe and orderly classroom.	❑ Creates a classroom culture that empowers students to collaborate.	❑ Empowers and encourages students to create and maintain a safe and supportive school and community environment.	

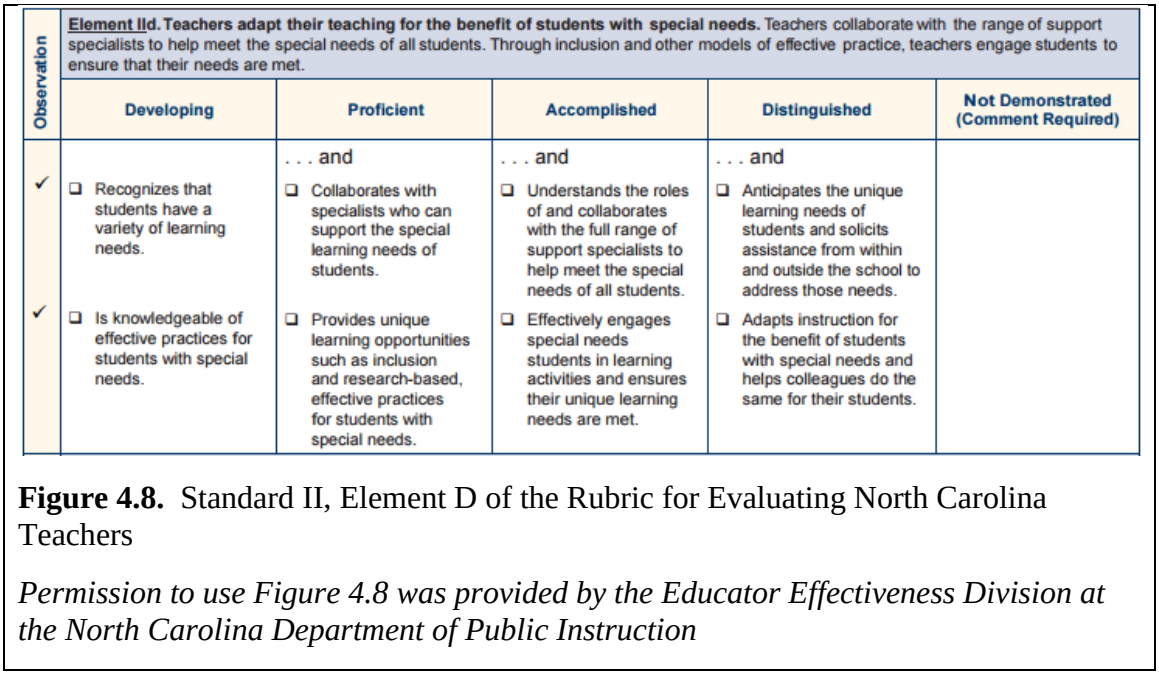
Figure 4.7. Standard I, Element A of the Rubric for Evaluating North Carolina Teachers

Permission to use Figure 4.7 was provided by the Educator Effectiveness Division at the North Carolina Department of Public Instruction

Selection of this element for the Fishbowl Intervention was based on several determining factors that emerged throughout Phase I. After viewing and evaluating SS1 on November 18, 2014, the OCT Pilot participants engaged in a debriefing conversation with the NCDPI NCEES Consultant. As the group reviewed Standard 1a, she reminded all OCT Pilot participants that only the check marked row of descriptors were observable during a classroom observation. She emphasized that the rating could not be "Developing" as this rating had no descriptor on the check marked row. However, 25% of Concordia participants had selected "Developing" as their rating for this element in SS1. During SS2, target agreement fell from 58% to 42%, and one Concordia participant selected "Developing" as their rating for Standard 1a. The design team selected Standard 1a due to the fact that this standard has five elements and 33 descriptors, yet only one

element with three descriptors is observable in the classroom. Even after two scoring studies, target agreement on the element declined, and at least one observer selected "Developing" for the target rating, although this rating is not possible based on the rubric. Based on these misconceptions and the decline in target proficiency from SS1 to SS2, the design team determined that Standard 1a was worthy of a deeper review during the Fishbowl intervention.

Standard 2d. Standard 2 addresses establishing a respectful environment for a diverse population of students (See Figure 4.8).



Observation	**Element IId. Teachers adapt their teaching for the benefit of students with special needs.** Teachers collaborate with the range of support specialists to help meet the special needs of all students. Through inclusion and other models of effective practice, teachers engage students to ensure that their needs are met.				
	Developing	**Proficient**	**Accomplished**	**Distinguished**	**Not Demonstrated (Comment Required)**
✓	❑ Recognizes that students have a variety of learning needs.	. . . and ❑ Collaborates with specialists who can support the special learning needs of students.	. . . and ❑ Understands the roles of and collaborates with the full range of support specialists to help meet the special needs of all students.	. . . and ❑ Anticipates the unique learning needs of students and solicits assistance from within and outside the school to address those needs.	
✓	❑ Is knowledgeable of effective practices for students with special needs.	❑ Provides unique learning opportunities such as inclusion and research-based, effective practices for students with special needs.	❑ Effectively engages special needs students in learning activities and ensures their unique learning needs are met.	❑ Adapts instruction for the benefit of students with special needs and helps colleagues do the same for their students.	

Figure 4.8. Standard II, Element D of the Rubric for Evaluating North Carolina Teachers

Permission to use Figure 4.8 was provided by the Educator Effectiveness Division at the North Carolina Department of Public Instruction

Standard 2d also proved to be problematic for the Concordia OCT Pilot participants throughout the study. During SS1, only 25% of participants agreed with the target score. Raw data from observable mini-lessons between SS1 and SS2 indicate that only 16.6% of Concordia participants who completed the mini-lessons agreed with the

target score. Moreover, in SS2, target agreement remained low at 33%. Although all eight of the descriptors of Standard 2d can be observed during a classroom observation, many of the descriptors require knowledge and understanding beyond the observation itself in order to make an accurate determination of the rating. Mr. Hines, principal of Concordia High School, shared with the group during a principals' meeting that he had heard the NCDPI NCEES Consultant say at the Fall Principal READY Meeting to a large group of principals representing 15 LEAs that evaluation should be based on what an evaluator sees and what the evaluator knows. Over the course of the OCT Pilot, participants complained that it is challenging to discern the level of collaboration in a video when the observer cannot converse with the teacher and no additional support personnel appear in the video. Mr. Hines explained to the group that effective and accurate rating of Standard 2d relies heavily on the collaboration and support from specialists that is almost impossible to determine from the classroom observation in isolation. This insight was important to the design team and provided the impetus for the group to select this element for review during the Fishbowl Intervention. Because the Fishbowl would take place at Southwest Elementary, Mrs. Felton, the principal, would be able to share this "outside the observation" information with observers. The design team was interested in whether or not observing in a real classroom would alter the participants' ability to evaluate this element accurately.

***Standard 4a*.** With eight elements and 50 descriptors, Standard 4 has the most observable elements and descriptors of all the standards. The magnitude of this standard made it one that the design team wanted to review during the Fishbowl Intervention. One

of the elements selected for review was Standard 4a. Standard 4 addresses pedagogy and how teachers facilitate instruction for their students (See Figure 4.9).

Observation	**Element IVa. Teachers know the ways in which learning takes place, and they know the appropriate levels of intellectual, physical, social, and emotional development of their students.** Teachers know how students think and learn. Teachers understand the influences that affect individual student learning (development, culture, language proficiency, etc.) and differentiate their instruction accordingly. Teachers keep abreast of evolving research about student learning. They adapt resources to address the strengths and weaknesses of their students.				
	Developing	**Proficient**	**Accomplished**	**Distinguished**	**Not Demonstrated (Comment Required)**
		. . . and	. . . and	. . . and	
✓	❑ Understands developmental levels of students and recognizes the need to differentiate instruction.	❑ Understands developmental levels of students and appropriately differentiates instruction.	❑ Identifies appropriate developmental levels of students and consistently and appropriately differentiates instruction.	❑ Encourages and guides colleagues to adapt instruction to align with students' developmental levels.	
✓		❑ Assesses resources needed to address strengths and weaknesses of students.	❑ Reviews and uses alternative resources or adapts existing resources to take advantage of student strengths or address weaknesses.	❑ Stays abreast of current research about student learning and emerging resources and encourages the school to adopt or adapt them for the benefit of all students.	

Figure 4.9. Standard IV, Element A of the Rubric for Evaluating North Carolina Teachers

Permission to use Figure 4.9 was provided by the Educator Effectiveness Division at the North Carolina Department of Public Instruction

The Concordia participants in the OCT Pilot struggled with Standard 4a since the summer 2014 mini-pilot. During the mini-pilot, none of the seven participants agreed with the target score. Data from SS1 revealed that only 58% of participants agreed with the target score. On February 4, 2015, the participants in the OCT Pilot reviewed the two OCT lessons on Standard 4a collaboratively during a principals' meeting. Each participant watched and scored the two lessons. After each lesson, the group engaged in facilitated discourse around the element and descriptors as well as the limitations of the video and platform. Feedback from participants included that the video angles, lack of

quality resources (such as lesson plans, student handouts, etc.), and brevity of these video lessons made evaluating the lessons very challenging. Only 29% of participants agreed with the target rating during the facilitated viewing of Standard 4a. Furthermore, SS2 results indicated that only 17% of participants agreed with the target score. This decline in target agreement precipitated a need to review this element during the Fishbowl Intervention.

Standard 4d. Standard 4d relates to how teachers integrate and utilize technology in their instruction (See Figure 4.10).

Observation	**Element IVd. Teachers integrate and utilize technology in their instruction.** Teachers know when and how to use technology to maximize student learning. Teachers help students use technology to learn content, think critically, solve problems, discern reliability, use information, communicate, innovate, and collaborate.				
	Developing	**Proficient**	**Accomplished**	**Distinguished**	**Not Demonstrated (Comment Required)**
✓	❑ Assesses effective types of technology to use for instruction.	. . . and ❑ Demonstrates knowledge of how to utilize technology in instruction.	. . . and ❑ Integrates technology with instruction to maximize student learning.	. . . and ❑ Provides evidence of student engagement in higher level thinking skills through the integration of technology.	

Figure 4.10. Standard IV, Element D of the Rubric for Evaluating North Carolina Teachers

Permission to use Figure 4.10 was provided by the Educator Effectiveness Division at the North Carolina Department of Public Instruction

Upon review of this element, it is important to note that the current North Carolina Professional Teaching Standards were revised between 2006 and 2008, and the North Carolina State Board of Education adopted the standards in 2008. Since 2008, the availability of classroom technology has increased exponentially; the resources, both virtual and physical, have expanded and changed in ways we could not have imagined

when the standards were written; and the expectation that districts, parents, and students have for the use of technology in schools has morphed with each technological advance.

The Concordia OCT Pilot participants reviewed lessons on Element 4d together on February 17, 2015. The participants were 55.5% in agreement with the target rating on the two videos in the element lesson on February 17. This was a slight improvement from the 50% agreement on SS1. Conversations continued to revolve around the differences in the expectation of technology use today and the expectation expressed in the rubric. The participants asked to view this element in a real classroom so that the group could get a closer look at how the teacher and students are utilizing technology to gain a clearer understanding of the demands of the standard and discuss "look-fors" together. Even though 67% Concordia OCT participants agreed with the target in SS2, the rich, ongoing conversations about this element during principals' meetings and after each scoring study prompted the design team to include this element in the Fishbowl Intervention.

Further considerations. Standard 3 was omitted completely due to the fact that in the target videos, the raw footage was taken from The MET Project prior to implementation of the Common Core State Standards. Since the footage was captured in the state of New York and not in North Carolina, the design team omitted a review of this standard. The group agreed that collecting data from video footage filmed in North Carolina where teachers purposefully attempt to align instruction with the standards would be more authentic and may yield more accurate data on Standard 3.

Analysis of Phase II results. The design team developed one goal to drive the work behind Phase II of the intervention. SS1 provided a baseline for participant target

agreement in November 2014. SS2 provided a benchmark assessment of participant target agreement in March 2015. The results of this assessment provided the implementation team with the opportunity to analyze the data to target specific elements for intensive focus. After a five-month study of standards, overall target accuracy increased by 5%. However, overall target accuracy on the four elements selected for deeper review dropped by over 12%.

On May 4, 2015, the participants completed SS3, a third full-length class lesson. Participants again rated all seventeen observable elements. To measure the effectiveness of the intervention, a paired samples t-test was administered to compare participants' performance on the four elements that were a focus in the intervention in both SS2 and in SS3. Following the administration of SS3, target agreement was analyzed and evaluated. Table 4.9 provides the analysis of that data.

Table 4.9 Comparison of Focus Elements on Scoring Study 2 and Scoring Study 3

	SS2		SS3						
	M	SD	M	SD	Difference	t(11)	p	95%CI	Cohen's d
Performance	35.41	29.11	60.41	24.90	25.00	-2.708	.020	[-45.3,-4.68]	2.60 {The formula for Cohen's d is M_1-M_2/√((s_1^2+s_2^2)/2))}

In order to understand the impact the Fishbowl Intervention had on our participants, we measured their target agreement on both SS2 and SS3. We wanted to understand if our participants demonstrated statistically significant improvement on the four elements that were selected as a focus for the intervention. The four elements were selected based on the collective low percentage of accuracy on the elements in SS2 (M=35.41, SD=29.11). It appears that our participants had a much higher rate of

accuracy on the four elements after the intervention as measured by SS3 (M=60.41, SD=24.90) (M=30.48, SD=21.23) with a mean difference of 25.00. Further analysis with a paired-samples T-test revealed the difference between performance on these selected elements from SS2 to SS3 was statistically significant, t(11)=-2.708, p=.020. Such findings are encouraging for educational leaders seeking to improve evaluation practices. It is also important to note that target discrepancy declined from 16.6% on the selected elements in SS2 to 0% on these four elements in SS3 as well.

Goal results. Phase II Goal: Concordia Public Schools evaluators and district leaders will participate in the Fishbowl Intervention so that target agreement will increase on the elements selected for the intervention by May 2015 as measured by a comparison of focus elements in Scoring Study 2 and Scoring Study 3. The paired samples t-test concluded that there was a statistically significant improvement at the $p<.02$. With a sample size of 12 participants, this result is particularly significant. Furthermore, no scorers were discrepant on any of the four elements that were a focus in the Fishbowl Intervention. The data suggest that a facilitated approach with prolonged engagement with the elements is successful in improving target agreement.

Identification of change concepts. Qualitative data coded from transcripts of the Fishbowl Intervention recording provide deeper insight into how participants process the act of evaluation and how they make meaning out of the standards. Several major themes emerged from the transcripts. These themes include the power of collective thought, the importance of understanding the language of the elements and identification of evidence, and the significance of the post-observation conference. These provided the team with

change concepts that mitigate the major driver for improving principal knowledge and skill in teacher evaluation.

Provide opportunities for collaborative learning. The power of the collective emerged as a significant theme in the discussion. Dr. Merta, the Concordia superintendent, who participated in the Fishbowl Intervention, shared, "…the collective thought is more powerful than the individual thought. Just like we're doing now. It adds validity and adds a different perspective when I have a specialist who says, 'Have you considered…?" As the conversation around what was observed in the classroom began to develop, another example of the importance of collective understanding emerged. Mr. Black shared, "I would say the ability for us to gather ideas and share what we've seen on the videos and stuff like that has been very beneficial. If you're sitting in your office and trying to do it by yourself, the distractions that were mentioned earlier or…just …it's through your lens only. You don't see stuff – other people's stuff." The Fishbowl Intervention improved collective competence of the group. Not only did target agreement improve, but no participant demonstrated rater discrepancy on any one of the four focus elements.

Improve skills in identifying evidence. Identifying evidence and the language of the elements also emerged as an important theme in the Fishbowl Discussion. One of the two groups engaged in a lengthy discussion around the language of the *Distinguished* descriptor in Element 4d that addressed student engagement. The members of the group discussed the difference between compliance and engagement, and ratings varied on this element. After several participants had pointed out that they thought there was no evidence all students were being successful in the class, the NCEES Consultant brought

them back to the language of the standard. She said, "There's nothing here [in this element] that says that students were successful, that all students were engaged, or all students were successful. It just says 'provides evidence of student engagement.'" The consultant's ability to refocus the group on the language of the element was a significant moment of clarity for the group. Refocusing on the standards and reading them multiple times to ensure our ratings are based on the language of the standard and the evidences that align with those particular ratings is imperative to high-quality, accurate evaluation. Two minutes later, as the conversation continued, Mrs. Higgins, the Director of Elementary Education, asked the group to refer to the language in Element 4d – *teachers help students use technology to learn content, think critically, solve problems, discern reliability, use information, communicate, innovate, and collaborate.* "I don't know that you can say that everything happened in that gray area for her to be a distinguished person. Where is the evidence…?" Once the NCEES Consultant set a precedence of leading the participants into the language of the elements to hone in on specifically what the standard demands, more and more participants referred to the language of the elements or finding evidence that aligned with the demands of the element.

Develop shared meanings of elements. Participants openly shared that often the language of the elements is confusing. During the discussion of Element 1a, one of the facilitators pointed out a misconception in the interpretation of the language of the element, "I'd like to point out the sole semantic here, that they're empowered to collaborate, not that they're empowered and collaborating." Mr. Thompson stated, "That's what gets me tripped up sometimes, is the verbiage." Throughout the 80 minute discussion, both facilitators continuously reminded the participants to re-read the

language of the descriptors and to pinpoint evidence from the observation to back up ratings. I found this practice as one that any good facilitator should use when working with standards, either during a collaborative walkthrough or any evaluation exercise.

Integrate information from the post-observation conference. One other important theme that emerged was the significance of the post-observation conference. When Mr. Hines shared that he made an assumption about the teacher to rate her accomplished when he only had evidence that she was proficient, the facilitator shared, "Again, I think the important take away here is not to make assumptions. The conversation is the important key. You'll only find these things through the conversation, but just in observation, we don't know these things." At one point, the superintendent shared, "After the conversation she may end up accomplished, depending on how the conversation goes. That's why the conversations is such a key part of this. The old system is just – you checked what you saw and moved on. The conversation matters here." While discussing Element 4a, Mr. Black said, "…unless you have that conversation with her, it's going to be hard to tell whether or not she's reviewed any of her resources." As the Fishbowl conversations deepened, the participants were not only identifying the language of the descriptors to share their evidences, but they were all making very important statements about how much information cannot be determined without the post-observation conference. Dr. Merta shared, "That's why the conversation is important. If I'm the teacher you are talking to, I would say, 'Here, let me show you what came from that,' or 'After you left, this is what…' I go in with what I think they've done, but the evidence can shift if further. It never shifts downward…" Mrs. Higgins also mentioned the importance of the post-observation conference, "I think it's very

crucial for us to have those conversations with the teachers up front and say to them, 'I'm going to see certain things. I'm not going to see everything. Understand that, and I want you to come back and have that conversation with me.'" The facilitators and all participants agreed that having the conversations with teachers following the observation is critical to accurate ratings. This is one practice that the OCT Pilot did not allow us to do. However, not having the ability to ask the teacher or students questions did force participants to refer back to the language of the elements and to the evidence that was available.

Fishbowl reflection. Following the Fishbowl Intervention, participants were provided with an anonymous survey regarding their experiences in the Fishbowl Intervention as well as the OCT Pilot and other evaluation exercises the group had experienced throughout the 2014-2015 school year. Of the 12 participants, 100% who participated in the Fishbowl Intervention perceived the activity to be beneficial. When asked "What was the most beneficial part of this exercise?,"100% of participants made mention of the importance of the discussion, particularly in terms of citing evidence and the language of the standards, and hearing others' points-of-view. All participants who participated in the Fishbowl Intervention agreed that the experience improved their practice as an evaluator. Likewise, all participants agreed that both participating in the OCT Pilot and watching the modules and discussing them during principals' meetings improved their practice as evaluators.

Conclusions. The comparative data from SS2 to SS3 on the four focus elements of the Fishbowl Intervention suggest that collaborative walkthroughs with explicit emphasis on the language of the standards, evidence, facilitated discussion, and

opportunities to both engage in conversation and observe conversation by other practitioners is an effective method for improving target accuracy. Based on a paired samples t-test, the participants demonstrated a statistically significant improvement at the $p<.02$ in target accuracy on the four focus elements from the Fishbowl Intervention from SS2 to SS3. Target accuracy improved from 35.4% to 60.4%.

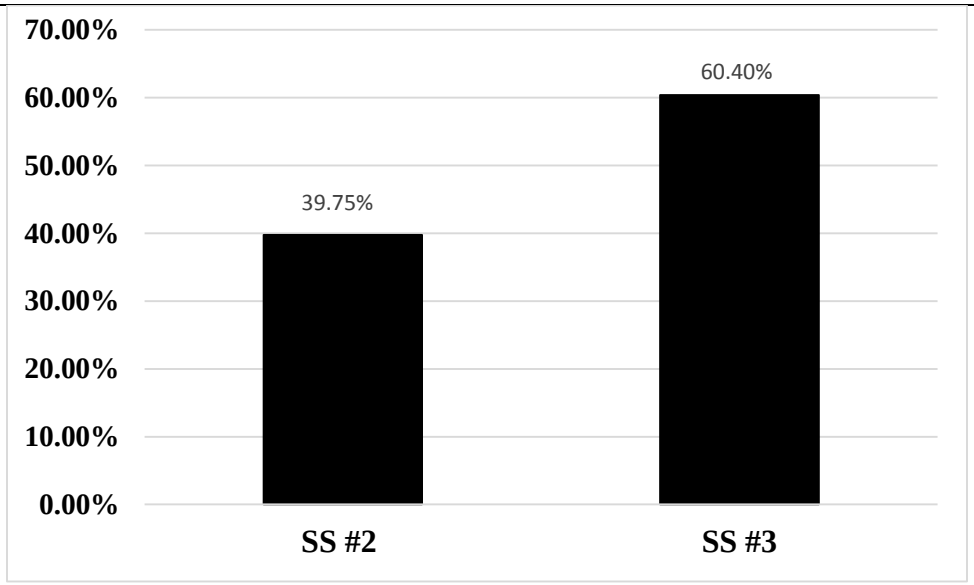


Figure 4.11. Comparison of target agreement on focus elements between Scoring Study 2 and Scoring Study 3

The intervention itself was multi-faceted and included opportunities for participants to engage with evaluation experts from NCDPI and provided opportunities to review the language of the elements, observe real classrooms collaboratively, engage in prolonged discourse inside the fishbowl, and listen to discussion around the elements as an observer from outside the fishbowl. Both performance and participant perception of the

intervention provided evidence that the Fishbowl Intervention was a successful intervention in improving target accuracy.

CHAPTER 5: IMPACT OF THE IMPROVEMENT PROJECT

The discrepancy between how evaluators rated teachers in one district and the concomitant student growth data provided the impetus for this improvement project. Over the course of approximately one year, evaluators in Concordia Public Schools engaged in a variety of evaluation improvement activities designed to improve target agreement with a pre-determined rating as well as reduce scoring bias and rater discrepancy. Data suggest that the project was successful in that target agreement improved throughout the project, and scoring bias and rater discrepancy decreased. The focus of this chapter is to report on the goal of the project and its impact on planning of future support in Concordia Public Schools.

Project Intervention Outcomes

Participation in the OCT Mini-Pilot, the OCT Pilot and the Fishbowl Intervention during the 2014-2015 school year provided principals and district leaders with opportunities for rich and meaningful conversations about the North Carolina Educator Evaluation System, the standards, descriptors, and elements of that system, and how to rate teachers effectively using the system. Qualitative and quantitative data documented incremental levels of improvement in the collaborative approach to implementing the OCT Pilot. Furthermore, a comparison of the four elements that were a focus of the Phase II Fishbowl Intervention assessed in SS2 and SS3 suggest that the interventions implemented in Concordia improved target agreement significantly. The data also suggest that the collaborative approach may have had more of a positive effect on improvement than simply completing the pilot alone. Quantitative data gathered at the

conclusion of the second phase were significant in developing the Phase III interventions to be implemented during the 2015-2016 school year.

Participant Perception of the Project

Qualitative data, collected from an anonymous survey provided to participants on April 28, 2015, indicated that 100% of participants agreed that participation in the OCT Pilot, the collaborative viewing and discussion sessions during principals' meetings, and the Fishbowl Intervention were all exercises that improved their practice as evaluators. However, participants did suggest, both in their reflection and in the Fishbowl discussion on April 28, that more support with coaching, post-observation conferencing, and providing feedback about how teachers can improve their practice were all areas for improvement.

During the discussion, the NCDPI Consultant asked participants what information they wanted her to take back to NCDPI, and one participant shared completing the OCT Pilot "as a group like we've done, it's been very beneficial." Another participant focused on the importance of the post-observation conference and how the pilot could be improved if NCDPI could find a way to incorporate this experience into the training. On the survey participants completed at the end of the Phase II intervention, the need for support with coaching, providing high-quality feedback to help teachers improve, and conducting the post-observation conference emerged as needs in Concordia.

The feedback provided by participants in the improvement project led the design team to incorporate coaching and conferencing professional development sessions for evaluators into the professional development plan for the 2015-2016 school year. These opportunities included support with coaching, conferencing, providing feedback, and

having difficult conversations with faculty and staff members. From the pre-implementation phase, the OCT Mini-Pilot, through Phase II, the Fishbowl Intervention, participants consistently provided the same feedback – they valued and wanted more time to collaborate and discuss ratings, to spend time in real classrooms and discuss instructional practices and alignment to standards, and to receive additional support with holding difficult conversations and coaching teachers during post-observation conferences. Each of these needs was explicitly addressed in Phase III – 2015-2016 Interventions.

Phase III: 2015-2016 Interventions

In theory, the design and implementation teams in Concordia Public Schools would plan the Phase III Implementation of this evaluator improvement project solely based on the copious amount of qualitative and quantitative data collected during the 2014-2015 school year throughout the previous phases of the project. However, unlike a clinical setting where the environment and participants can be controlled in some aspects, Concordia experienced significant personnel changes between the 2014-2015 school year and the 2015-2016 school year that change the trajectory of the planning and implementation of Phase III.

Inconsistent Impact Due to Turnover

When the project began, the Phase III Cycle was slated to be a continuation of the work completed in the Pre-Phase Cycle and both Phases I and II. However, due to the a great deal of turnover in Concordia's site-based and district-based leadership between the 2014-15 and 2015-16 school years, the design team re-evaluated the Phase III interventions. Fifty percent of site-based leadership changed at the end of the Phase II

Fishbowl Intervention, and three of the five district-level participants in the project left the district in the summer of 2015. With so many new leaders, the Phase III Intervention was designed more around lessons learned and developing a foundation for high-quality evaluation rather than developing a next step for cohort improvement. The design team took into account the overarching feedback and data from Phase I, OCT Pilot, and Phase II, Fishbowl Intervention, as well as much of the informal, ongoing discussions regarding evaluator needs. The PDSA Cycle in Figure 5.1 provides the key components of the Phase III iteration of the implementation plan.

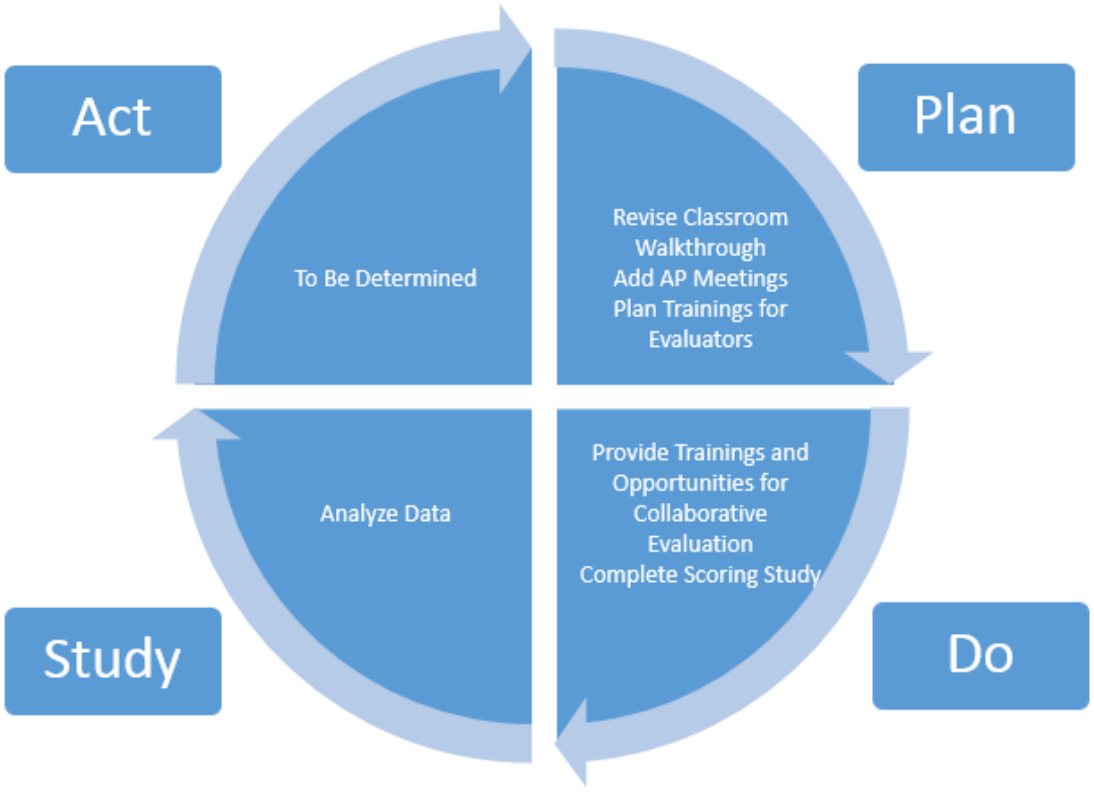


Figure 5.1. Phase III PDSA cycle - 2015-2016 interventions implementation

Figure 5.1. The Phase II PDSA cycle implementation was devised based on data and feedback from the Pre-Implementation Phase (OCT Mini-Pilot), Phase I (OCT Pilot), and Phase II (Fishbowl Intervention). This iteration of the project will span the 2015-2016 school year.

Takeaway for leaders. Although the data collected during the improvement project strongly suggested that evaluators in Concordia improved their evaluation practices, and their feedback reinforced that they valued and desired more support with evaluation, the turnover in leadership introduced a challenging problem for the design and implementation teams. The question became – how can we continue to build upon the skills of the evaluators who have benefitted from this project while meeting the needs

of the new evaluators who may or may not have the prerequisite skills we now assume our returning evaluators have? This question drove our decisions for implementation of Phase III. After taking into consideration the data and the changes in personnel, we selected interventions for the 2015-2016 school year. These interventions will move the returning evaluators forward and meet their self-reported needs while meeting the needs of the new evaluators. This plan will require providing professional development for novice evaluators in a manner not previously explored in Concordia.

Preparation for the 2015-2016 School Year

One impact of the project related to plans for continuing professional development for all principals and assistant principals. The Phase III plan was developed based on the formal and informal feedback provided by participants throughout the pre-implementation phase, Phase I, and Phase II of the project. Quantitative data demonstrated that the interventions implemented during the 2014-15 school year were successful in improving evaluator effectiveness in Concordia Public Schools as participants continued to show improvement in target accuracy, scoring bias and rater discrepancy on each iteration of the project. The greatest gains took place in Phase II – Fishbowl Intervention, as participants grew from 35.4% rating accuracy on four elements in SS2 to 60.4% target accuracy in SS3. An analysis of qualitative data indicated that participants believed the interventions to be successful because they had an opportunity to visit classrooms in small groups and engage in sustained discourse regarding the language of the elements of the Rubric for Evaluating North Carolina Teachers as related to the classroom observations. Principals who participated expressed a need to provide assistant principals with opportunities to engage in evaluation improvement exercises, to

receive support with providing feedback to teachers, coaching teachers effectively, and engaging in difficult conversations regarding evaluation. This feedback played a vital role in the development of Phase III.

Integrating coaching training to improve evaluation practices. In response to feedback from participants in Phases I and II, the design team developed a plan to provide two layers of professional development to district and school leaders. First, during the summer of 2015, the CAO, who is a certified *Crucial Conversations* trainer, provided the VitalSmarts *Crucial Conversations* training to district and school leaders. *Crucial Conversations* training provides participants with the skills to hold difficult conversations in a way that promotes dialogue and discussion. The training provides support with recognizing when an individual is becoming defensive or is withdrawing, and then arming them with the skills they need to guide the conversation in a way that brings about results. The purpose of the training was to provide an overarching layer of support for how to hold crucial or difficult conversations, such as those that may follow an observation. A second training will be provided in spring 2016 for any leaders who could not attend.

In order to hone in on the specific need for how to provide feedback and coach teachers effectively, the next layer of support for post-observation conferences will be provided by the NCDPI NCEES Consultant. She has agreed to provide *Coaching Conversations* training to all administrators in Concordia during the 2015-16 school year. This training, developed by the NCDPI NCEES Consultant, was suggested by principals who attended the NCDPI-sponsored Principal READY Meeting in the spring of 2015 and shared that this training is a needed next step in improving evaluation practices in the

district. *Coaching Conversations*, unlike *Crucial Conversations*, is designed to provide principals with the skills to coach teachers and support them in the endeavor of improving teacher practice. While *Coaching Conversations* does reiterate many of the constructs of *Crucial Conversations,* the NCDPI training is directed toward principals and their work with teachers whereas the VitalSmarts *Crucial Conversations* training provides foundational support for conducting difficult conversations in any arena, including the workplace, in personal relationships, or with complete strangers. These two trainings coupled together will provide leaders in Concordia with skills and understandings to improve evaluation practices through successfully engaging in difficult conversations regarding teacher practice during post-observation conferences. Ultimately, by preparing principals to have more meaningful and non-confrontational conversations, we can develop more knowledgeable, skilled teachers who will develop classrooms where all students learn and grow.

Addition of AP meetings to support improved evaluation practices. The need for providing opportunities for assistant principals to improve their practice by participating in activities such as the OCT Pilot and collaborative observations emerged throughout a series of informal discussions during the 2014-2015 school year. Concordia has only five assistant principals – two at Concordia High School, two at Concordia Middle School, and one at Concordia Academy. Because the district has so few APs, there has been no formal structure for ongoing meetings with these leaders. One focus of these meetings will be to support the APs'work with the Rubric for Evaluating North Carolina Teachers and coach them on how to use the tool effectively as well as how to engage in effective post-observation conferences. Furthermore, OCT Pilot videos and

classroom visits will be a part of monthly AP Meetings. Initial elements for review will be selected from elements on which Concordia participants in the improvement project scored particularly low. Because three of the five assistant principals were hired in the summer of 2015, and none of these hires has administrative experience, only two of the APs have any evaluation experience outside their practicum experiences. Therefore, these types of collaborative experiences to improve the quality of evaluation are imperative to building capacity. Since the principals and APs will experience the same types of evaluation trainings and engage in similar conversations, our goal is to develop a shared language among evaluators as well as encourage the practices of reviewing the language of the standards, ensuring evidence exists to support ratings, and eliminating bias from the evaluation process.

APs will also be included in the *Crucial Conversations* and *Coaching Conversations* trainings. The goal of including APs as well as instructional coaches and district leadership is to develop a shared framework for how to initiate and successfully conduct a difficult conversation with a faculty or staff member. Ensuring that all leaders receive the same trainings and support will not only ensure consistency but also serve to prepare APs and instructional coaches for potential leadership roles they may accept in the future.

Redesigning the classroom walkthrough tool. For the past two years, the Concordia classroom walkthrough tool has been developed and revised to reflect district priorities that are aligned with North Carolina State Board of Education Goals (Appendix O). One of the goals of the CWT is to collect holistic data on pedagogical practices

throughout the district. The CWT includes a variety of "look-fors" that are priority for the district.

Construction and purposes of the CWT. The CWT was created in a Google Form, and all information is submitted electronically. The data can be displayed in a variety of ways. District and school leadership use this data to look at classroom trends, select appropriate professional development, and pinpoint instructional expertise and needs. Furthermore, our CWT includes a script that ensures that the moment a CWT is submitted, the teacher being observed receives feedback on his walkthrough. The teacher can review each element of the walkthrough and see what the evaluator observed as well as anecdotal comments the observer may have provided. The CWT, however, is non-evaluative. It is used solely as an indication of what teaching and learning "looks like" in Concordia, to provide non-evaluative feedback to teachers, and to align evaluation practices.

Evaluators across the district are required to complete a minimum of five walkthroughs each week. Walkthroughs should be completed with another evaluator either from the school or district level. The goal of completing collaborative walkthroughs is to discuss "look-fors" and ensure all evaluators have a common understanding of pedagogical practices. Furthermore, evaluators can pinpoint areas for improvement or best practice to share either with the school faculty or across the district.

Changes to the 2015-2016 CWT. In previous school years, all fields in the CWT have been required. However, based on the statistically significant improvement from SS2 to SS3 on target agreement when only a few elements of the NCEES Rubric were addressed, the design team, along with the district's executive leadership team and

principals, determined that providing the option for focusing only on specific elements rather than the entire set of elements may provide opportunities for more targeted feedback and collaborative conversations. The data from Phase II was the most significant factor in this change in how the CWT could be used. The walkthrough revisions were completed in August 2015, and the modified tool was available for district use on October 1, 2015.

Principals' meetings. During the past two years, principals' meetings were held in schools. At each of these meetings, time was allocated for a focus on evaluation. During some meetings, principals and district leaders engaged in classroom walkthroughs and discussed the "look-fors" on the district's CWT document. Most of the 2014-2015 school year was devoted to the OCT Pilot videos and engaging in facilitated discussion tailored around the specific elements in each module. Based on the results from both Phases I and II of the improvement plan, Concordia will continue a focus on evaluation during each of the principals' meetings of the 2015-2016 school year. As described earlier in this chapter, collaborative walkthroughs will now focus on only a few elements of our district CWT document or on one or two elements of the Rubric for Evaluating North Carolina Teachers. To continue to move forward in meeting the needs of principals, collaborative classroom walkthroughs will also use a newly-developed protocol that will include what coaching would look like.

Evaluating Success

One critical way in which this project influenced the principal training initiative going forward was in our development and use of improvement science methods to inform change projects. Our goal is to continue to provide support so that new and

veteran principals and assistant principals will continue to improve as evaluators, ratings will be more accurate, and teachers will receive quality coaching feedback. Our plan is to implement the strategies from this project that led to improvement in participant knowledge and skills related to teacher evaluation during the 2015-16 school year, and then give evaluators an assessment to determine what the strengths and areas for improvement are in the group.

Measurement of increments of improvement will be essential for this next phase. The OCT Pilot was equipped with a fourth full length video that was vetted and rated by master raters. Concordia participants did not view or rate this video. Our current plan is to coordinate with the NCDPI NCEES Consultant and request access to it so that we can use this video to measure rater accuracy, agreement, and scoring bias at the end of the 2015-16 school year. This quantitative measure, along with informal discussions, surveys, and classroom walkthrough data, should provide district leaders with the information needed to provide targeted support to evaluators in the district. The goal of this support is to provide teachers with the high-quality feedback they need to improve teaching and learning in their classroom in order to support student growth and achievement. Figure 5.2 provides a visualization of the goal of Phase III, a simplified version of the adaptation of "The Ripple Effect."

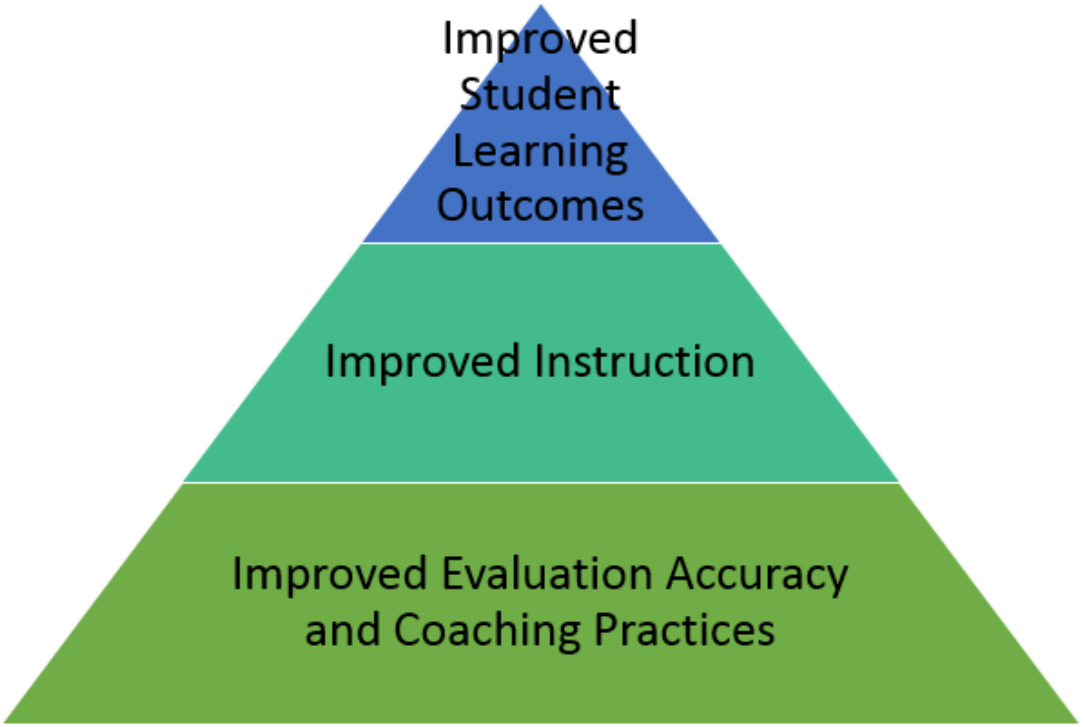


Figure 5.2. Visualization of Phase III Implementation Goal

This project was developed around an adaptation of "The Ripple Effect," the idea that if district leadership and state leadership worked collaboratively and synchronously to improve principal evaluation practices, that teacher effectiveness and student learning outcomes would improve. Although the data indicate that principal evaluation practices did improve in the controlled setting of the project, data collected from principal evaluations of teachers and state EVAAS data reveal that more work must be done in the area of evaluation improvement both at the state and district levels.

Overarching Project Goal Results and Conclusions

Overarching Project Goal: Concordia Public Schools principals will participate in the 2014-15 OCT Pilot and subsequent interventions so that the correlation between principal ratings on Standards 1-5 of the NCEES Rubric and Standard 6 will increase by October 2015 as measured by a comparison of EVAAS data from the 2013-14 school year and the 2014-15 school year. The impetus for this improvement project was the review of two years of correlational data that indicated little to no correlation between the way principals evaluated teachers and student growth as measured by the EVAAS value-added model. In November 2015, the NCDPI released the 2014-15 EVAAS data to districts for review. Again, I conducted the same Spearman-R correlational analysis on this data as I did in 2014 on the previous year's data. Table 5.1 provides the results of the data review

Table 5.1

2014-2015 Correlations between Assigned Teacher Ratings and Value-Added Growth Ratings in Concordia Public Schools

	St. 1	St. 2	St. 3	St. 4	St. 5	St. 6
Standard 1	1.00	.445**	.505**	.501**	.495**	.188
Standard 2	.445**	1.00	.420**	.521**	.383**	.208*
Standard 3	.505**	.420**	1.00	.570**	.523**	.145
Standard 4	.501**	.521**	.570**	1.00	.548**	.105
Standard 5	.219*	.495**	.383**	.523**	1.00	.219*
Standard 6	.188	.208*	.145	.105	.219*	1.00

*Correlation is significant at the 0.05 level (two-tailed)
**Correlation is significant at the 0.01 level (two-tailed)

In a side-by-side comparison, data indicate that alignment between principal ratings and EVAAS value-added growth data (Standard 6) improved from 2013-14 to

2014-15 on Standard 1, Standard 2, and Standard 5. Correlation between Standard 6 and Standards 3 and 4 declined from the 2013-2014 school year to the 2014-2015 school year. It is important to note that Standards 2-4 are representative of classroom instruction. Correlation between and among each of the standards 1-5 remained statistically significant at the 0.01 level (two-tailed). The lack of significant correlation between Standard 6 and the other five standards reveals yet a deeper issue that educational leaders face today. Data from this project suggest that the participants improved their ability to evaluate teachers in terms of target agreement, scoring bias, and rater discrepancy. However, no significant changes were evident in a side-by-side comparison of evaluation data from 2014 to 2015 in Concordia Public Schools. Table 5.2 provides a side-by-side comparison of data from the evaluations of all CPS teachers on the full observation cycle, those who received ratings on all five standards during both the 2013-2014 and 2014-2015 school years.

Table 5.2

2013-2014 / 2014-2015 Correlations between Assigned Teacher Ratings and Value-Added Growth Ratings in Concordia Public Schools

	2013-2014	2014-2015
Standard 1	.165	.188
Standard 2	.115	.208*
Standard 3	.227*	.145
Standard 4	.199*	.105
Standard 5	.113	.219*

*Correlation is significant at the 0.05 level (two-tailed)
**Correlation is significant at the 0.01 level (two-tailed)

Because no significant changes were noted in the formal CPS evaluation data, I suggest that understanding the evaluation tool itself may be only one of the problems

principals and other evaluators face. First, criticism of the use of value-added data is becoming more common (Darling-Hammond, 2015). Furthermore, the reauthorization of the Elementary and Secondary Education Act (ESEA) as the Every Student Succeeds Act (ESSA) may eliminate value-added models from use in teacher evaluation. Even though this action may solve a controversial issue regarding evaluation, the challenge of dealing with conflict as a leader still exists. When we look past all the improvements that CPS participants made during the project, one of the most important keys to success, I surmise, was the fact that none of the participants had to do any coaching or true evaluation of the teachers we rated. This project did not take into account the post-observation conference, the action steps for improvement, possible modifications to the teacher's PDP, ongoing follow-up and support for teachers who are underperforming, or the emotional responses that are generally a by-product of difficult conversations. These elements of the evaluation experience – the more qualitative, unpredictable, or individualized – cannot be accounted for in this project. Therefore, leaders in CPS must emphasize how to support evaluators with these elements of evaluation, not addressed in this project but to take center stage in the future.

Skilled evaluators affect instructional improvement through a deep understanding of the process of evaluation coupled with the ability to articulate constructive feedback to teachers effectively. As principals and assistant principals improve their ability to evaluate accurately and provide clear, specific coaching feedback to teachers in a caring and non-confrontational manner, teachers improve their skills in making informed instructional decisions to improve the quality of teaching and learning in their classrooms. Ultimately, the power of knowledgeable, skilled teachers to create learning

environments where all students thrive will be the testament to the impact of this and succeeding projects aimed at improving teacher evaluation.

CHAPTER 6: IMPLICATIONS FOR STAKEHOLDERS

The process of addressing an organizational problem in education requires a Networked Improvement Community (NIC). The foundation of improvement science is built upon learning from the NIC to develop a deep understanding of the problem, the system(s) that contributed to the problem, and a shared theory on how to address the problem incrementally (Our Ideas, 2016). Valuing the fundamental knowledge of educators, improvement science allows practitioners to engage in deep and meaningful work that can transfer out to the NIC to foster future improvements in other contexts. Although von Hippel (2005) posits that the work of innovative problem solving depends on highly localized problems that are contingent upon the context, the problem of practice central to this project has been discussed at the national level as *The Widget Effect*, at the state level through the research of Batton et al. (2012) and Tomberlin (2014), and at the local level through data analyses by Mullins (2015). At each level, the same problem of practice exists – teachers are rated as good or great, and the most effective teachers, according to achievement and growth, are rated the same as the most ineffective teachers. This project has allowed me to work with an NIC of leaders and practitioners from my own school district, other school districts, the NCDPI, and national consultants from Bloomboard and Empirical Education to improve evaluation in my own district as well as provide data and knowledge to the NIC for future improvements at the local, state, and national levels. This chapter relates some of the most significant implications, limitations, and recommendations from the improvement project.

Local Implications of the Project

This project began with the Adaptation of *The Ripple Effect* as a framework to visually depict the importance of the state-district partnership as the vehicle to co-develop the set of interventions to improve evaluation practices. According to Leithwood, Seashore, Anderson, & Wahlstrom (2004), "Leadership is second only to classroom instruction among all school-related factors that contribute to what students learn at school" (p. 7). Understanding that state leaders and district leaders influence principal practice that can lead to improved teacher quality was the impetus for this project. Participation in the pre-implementation phases, the OCT Pilot, and the Fishbowl Intervention all provided data to suggest that when district leaders and state leaders collaborate to support principals with evaluation, overall rating accuracy improves, and scoring bias and scoring discrepancy decrease.

Even though the framework encompasses evaluation, teacher quality, instructional feedback, and instructional quality, the project focused mainly on the component of evaluation.

Collaborative Learning Matters

Throughout the OCT Mini-Pilot, the OCT Pilot, and the Fishbowl Intervention, participants engaged in ongoing discourse that brought misconceptions to light and provided the opportunity to raise and answer questions about the evaluation rubric and evidences. Data from all three scoring studies indicated that target agreement improved and both scoring bias and rater discrepancy decreased over the course of the project.

This project and the interventions implemented throughout the course of the project suggest that the participants desire ongoing support with evaluation. They

consistently asked for more opportunities to engage in collaborative walkthroughs and evaluation exercises in future meetings and professional development opportunities. They indicated a need for more support with understanding the language of the standards and providing evidence. Concordia will continue focusing on these elements of evaluation during the 2015-2016 school year. In revisiting Figure 2.1 from Chapter 2, it is important to note that the emphasis on providing professional development and support to principals regarding evaluation is crucial to teacher and student growth.

Although the principalship can seem isolated, research suggests that principal practice improves when principals work collaboratively with their peers. According to Coffin and Leithwood (2000), "participating with others in authentic, non-routine activities" promotes on-the-job learning (p. 21). When principals collaborate around real problems and issues that they address daily, principals are "exposed to a broader, perhaps richer, palette of ideas and approaches" (Peterson & Cosner, 2005, p. 30). Our data suggest that when state leaders and district leaders work together to support principal evaluation practice, that practice does, in fact, improve. Now that we have seen results, Phase III will move to widen the scope of impact from just evaluation to improving teacher quality through evaluation. One of the most important implications for future interventions is that the participants in the project asked specifically for support with holding coaching conversations and providing meaningful, high-quality feedback to teachers. This desire to improve their practice through an emphasis on instructional quality directly aligns with the theoretical framework and serves as a logical next step in the cycle of continuous improvement. Continued work on evaluation practices coupled

with support with instructional feedback will move the district closer to improved student learning outcomes.

The Importance of Prolonged Engagement

The OCT Pilot provided participants with one form of prolonged engagement. The review of quantitative data from the three scoring studies in this project suggests that prolonged engagement in evaluation practices improved participants' practice on the scoring studies over time as participants' accuracy improved, and their propensity for bias and discrepancy declined. The lack of a body of research regarding the effect of sustained professional development for principals reinforces the importance of this study. However, many researchers have reported the effects of ongoing, prolonged professional development on educators in general.

According to Guskey and Yoon (2009), "effective professional development requires considerable time, and that time must be organized, carefully structured, purposefully directed, and focused…"(p. 497). Participants engaged in individual or group professional learning throughout the 2014-15 school year. Each month, the participants engaged in two separate evaluation improvement sessions that were between 30-120 minutes in duration. During some sessions, participants viewed OCT Pilot videos or scoring studies, rated the teachers in the videos, and engaged in collaborative discussions about the ratings. In other sessions, they engaged in collaborative classroom walkthroughs and discussions regarding evaluation. Between each of these meetings, participants watched four classroom videos in the OCT, rated the teachers in the videos, and received immediate feedback on their performance. Yoon et al. (2007) also reported that, "Studies with more than 14 hours of professional development showed a positive

and significant effect on student achievement from the professional development experience. Although this finding was related to teacher professional development, in this case, the data suggest the finding might relate to evaluators as well. During the course of this project, participants engaged in over 16 hours of professional development on improving evaluation practice.

Participants also experienced a different form of prolonged engagement during the Phase III Fishbowl Intervention when they spent almost three hours reviewing four elements of the NCEES Rubric, observing those four elements in a classroom, and then engaging in a discussion of those elements as well as a listening exercise where they observed another group discuss their observations as they pertained to the element in a separate classroom observation. Data suggest that this intervention had a statistically significant impact on target accuracy. For a district leader who is ultimately responsible for planning the content of monthly principals' meetings, having data to support the effect of prolonged engagement on practice is important to planning for future interventions.

Need for State-Level Support

Another observation that emerged from the project is that it is challenging in a small school district to evaluate the quality of evaluation without valid tools. Although district-initiated classroom walkthroughs and discussions centered on the elements and standards of the Rubric for Evaluating North Carolina Teachers have helped us with rater agreement and a deeper collective understanding of the standards, the state-initiated OCT Pilot provided master ratings to ensure our collective understanding was accurate. Because the master ratings of the scoring studies and the lessons in the OCT were scored

by expert raters, and the reliability and evidence of validity were approved by a respected psychometrician, meaningful conclusions could be drawn from the data to determine next steps.

Challenge of True Data-Driven Decision Making for Evaluating Professional Development

In Concordia, we do not have easy access to high-quality, reliable instruments by which to assess the impact of interventions. Many districts can hire companies or universities to develop tools and evaluate practices, but these groups are often expensive, and without grant funding, small districts like Concordia find it difficult to finance these services. There is a need for more support for districts that would like training around developing high-quality reliable tools with evidence of validity and the skills to use inferential and descriptive statistics to analyze the data to draw meaningful and accurate conclusions. Although doctoral programs provide practitioners with training and coursework in these areas, most of our jobs are so focused on the implementation of state initiatives; working with leaders, principals, and teachers to improve practice; and meeting deadlines that there is little time for the deep and time-consuming work of developing tools to measure success. During the course of this project, our design and implementation teams found that Improvement Science offers a set of strategies that can be customized to meet the needs of leaders as they address local problems, thus broadening the opportunities for even small districts to initiate data-driven improvements.

Subjectivity of Evaluation

One final observation is that evaluation is never an exact science, and the work of improving the quality of evaluation is never complete. Evaluators bring myriad biases

and a great deal of subjectivity based on their experiences as a student, a teacher, and an administrator into a classroom observation. Because districts are constantly hiring new administrators from outside the district as well as promoting leaders within the district, the process of working to improve the quality of evaluation must be ongoing. As we begin the 2015-2016 school year, our district has two new principals and three new assistant principals. As leaders come and go, processes must be put in place to maintain a constant focus on the importance of high-quality and accurate evaluation.

At the beginning of this improvement project the implementation team regarded target agreement as one of the most significant factors to help us identify success. However, after engaging in numerous conversations regarding the subjective nature of the ratings and the quality examples of evidence provided by participants regarding the ratings, the team determined that target discrepancy was a far more reliable measure of whether or not each iteration was a success. Because evaluation is indeed subjective, the argument can be made, in many cases, to support the determination of a rating one level below or one level above the target. However, ratings two or more levels above or below the target reveal a misconception or lack of true understanding of the standard, elements, or descriptors provided. This realization helped us redefine whether or not the project was successful.

The concept of evaluation is massive and messy. In North Carolina, evaluators must rate teachers on 147 elements on their summative evaluations. Of these 147 elements, 133 are observable during a classroom observation. Because the Rubric for Evaluating North Carolina Teachers is so detailed, district leaders must constantly

provide guided support on the language of the elements, "look-fors," and evidences of the elements in practice.

Principals also need more opportunities to engage in professional development developed by or facilitated by evaluation experts. In Concordia, this aspect of the project brought a sense of formality and importance to the experience. Furthermore, when participants disagreed with the ratings, having a validated score with explanation for each OCT Pilot video rating and having an evaluation expert in the room, in the Fishbowl discussion, was powerful in ensuring buy-in and acceptance of ratings that were not originally accepted as accurate. Although the NCDPI does not have the capacity at this time to work individually with every district in North Carolina, restructuring regional support to provide access to state-trained evaluation experts to districts across the state is essential.

Next Steps

Our next steps include focusing this work on a continued emphasis on evaluation. A deep analysis of all the data in this project reveals that although our evaluators grew significantly in the areas measured by the scoring studies used in the project, there is still a great deal of work to be done to ensure this isolated learning transfers to practice in ratings principals assign to teachers on their ongoing observations and summative evaluations. Continued emphasis will be on the language of the standards, evidences, and perceptions through classroom walkthroughs and professional development on individual elements of the Rubric for Evaluating North Carolina Teachers.

Implementation and Analysis of Phase III

Concordia will implement Phase III of this project over the 2015-2016 school year. Based on the conclusions we have drawn from Phases I and II regarding the importance of collaborative learning and ongoing support, our next step will be to determine whether this iteration of the improvement plan yields more positive results Furthermore, this phase will be used to complete an additional PDSA cycle to determine subsequent phases for improvement.

Although the measure of overarching success will be determined by the correlation of principal summative ratings of teachers on Standards 1-5 in the NCEES and Standard 6, the design team needs to work together to develop a plan for more informal and formative evaluation of the intervention. The OCT has one more full-length video that we could use as Scoring Study 4 (SS4) with permission from the NCDPI. One next step will be to contact the NCDPI to determine whether or not the Bloomboard platform will be available and if we will be able to use SS4 as a final evaluation of the project and receive the raw data from Empirical Education to run an analysis when Phase III interventions are complete.

Professional Development on Coaching and Conferencing

Based on the qualitative feedback we received from principals, we will also partner with the NCDPI to provide professional development and support for principals on the topics of providing feedback and coaching to classroom teachers. The NCDPI has created two support documents that we will use this year in Concordia to support our ongoing professional development. The first document is "NCEES: Questions for Post-Observation Conference and Summative Evaluation" (Appendix P). Because evaluators

have asked specifically for support with feedback and coaching, the curriculum team will use this document and to facilitate post-CWT discussions. The second document, "Evidences for Professional Teacher Standards 1-5," was compiled by the Educator Effectiveness Division at the NCDPI after being developed by North Carolina principals during the 2013-2014 Principal READY meetings (Appendix Q). The intention is to use these additional support documents to help clarify misconceptions as well as support principals with questions to ask during post-observation conferences and coaching sessions with teachers. Additionally, we are coupling this support with two trainings for evaluators – *Crucial Conversations* training and *Coaching Conversations* training. These trainings will equip evaluators with tools and strategies for engaging in challenging and often uncomfortable discussions with teachers.

Information for Use by Others in the Networked Improvement Community

Accurate, meaningful evaluation is an issue that we face not only in Concordia or in North Carolina, but also throughout our nation as a whole (Batton et al., 2012; Tomberlin, 2014; & Weisburg et al.,2009). This project is significant due to the vast networked improvement community struggling to improve the same system within their countless organizations. According to Bryk, Gomez, & Grunow (2011), in order for the "NIC to make headway towards constructive improvements on a complex problem, the community needs to detail the contours of its problem-solution space" (p. 15). This project yielded a great deal of important data and overarching conceptual ideas to consider for the NIC, including the impact of collaborative discussions on participants; the chunking of the professional development over time; and the need to include all possible evaluators, including assistant principals and district leaders, in the trainings to

improve the collective competence and understanding of the evaluation process and instruments. However, there are also many important aspects of the project that affected CPS. Each organization is unique, and members of the NIC will encounter their own considerations as they approach implementing improvements to evaluation practices. Before adapting this project and attempting their own iterations of the work, reviewing local implications of small sample size, participation and attendance, and attrition may prove helpful.

Small Sample Size

Having a small sample size (n=12) is an important consideration of this project. With a group of 12 participants, facilitators can hear from all stakeholders and make all aspects of the professional development meaningful and personal. Although not all districts have the luxury of working with a small group of participants when embarking on an improvement project, the strategies we used can help inform other districts and state agencies in the NIC who are working toward improving evaluation practices. However, having a representative group of only 12 is also a limitation to the findings of the project. With only 12 participants, the results may not be applicable to larger districts who have exponentially more principals to serve. Furthermore, some researchers or practitioners may not consider the project data significant due to the limited number of participants. Repeating this study in a larger district may yield more generalizable results and will add to the strategies for consideration in the NIC.

Participation and Attendance

Another limitation of the study is the fact that not all participants participated in all aspects of the study. Of the 12 participants, only seven participated in the OCT Mini-

Pilot. Although SS1, SS2, and SS3 were supposed to be completed in a controlled setting, only nine of the 12 participants completed SS1 in the controlled setting during the principals' meeting on November 18, 2014. During the OCT Pilot, nine of the 12 participants completed all 17 modules (34 lessons). One participant completed nine modules, one completed 10, and one completed 16. Only seven of the 12 participants were present for the collaborative viewing of SS2 in a controlled environment. The dynamic is typical of most school districts where conflicting priorities often result in impromptu changes in principals' schedules. Nevertheless, after a review of the results of SS2, it was clear that participants who were present on March 25 performed 11.2% better than those who completed SS2 independently. Furthermore, Table 6.1 provides data from the participants in the study who were present for SS1 but absent for SS2 and completed that scoring study independently.

Table 6.1 Comparison of Target Agreement Performance of Participants Who Completed Scoring Study 1 in a Controlled Environment but Completed Scoring Study 2 in an Independent Environment

	SS1		SS2						
	M	SD	M	SD	Difference	t(3)	p	95%CI	Cohen's d
Agreement	51.50	16.52	42.75	19.39	-8.75	2.376	.098	[-2.97, 20.47]	2.60 {The formula for Cohen's d is $M_1-M_2/\sqrt{((s_1^2+s_2^2)/2)}$}

This finding suggests that providing a controlled environment contributes to better performance. It appears that participants who completed SS1 in a controlled environment performed better on SS1 (M=51.5, SD-16.52) than they did on SS2, which they completed in an independent environment (M=42.75, SD=19.39) with a mean difference of -8.75. Further analysis of the paired samples T-test revealed that participants who completed SS1 in the collaborative setting but completed SS2 in an independent setting demonstrated a statistically significant decline in their performance, t(3)=2.376, p=.098. These data are important to consider for educational leaders planning professional development for administrators.

During the project, norms and expectations were provided for collaborative viewing sessions. Distractions were removed as participants were instructed to put their computers and phones to the side. Support documents and components from the Rubric for Evaluating North Carolina Teachers were provided, and participants gave their full attention to the exercise. The same four participants who were present for SS1 and absent from SS2 were present for the administration of SS3 in the controlled environment. This group improved their performance from 42.7% target agreement to 54.5% agreement.

Attrition

All participants completed through Phase II of the project. However, one participant was diagnosed with a terminal illness before the study ended and passed away one week after SS3. The score of this participant on SS3 was 35% lower than SS2. Because of the small sample size, however, I did decide to keep the participant in the study. Two participants retired after Phase II, and one participant moved to another district. With only eight of the 12 initial participants remaining, the intervention will now be tailored not only to the original participants but also to the four new members of our administrative team. This is a common phenomenon in districts. The goal is to continue to grow leaders in their understanding of evaluation and their ability to evaluate effectively.

Recommendations for District-Level Leaders

District leaders who work with and evaluate principals must connect evaluation results to authentic professional learning. District leadership must provide support and guidance for principals in the form of "professional learning focused on instructional leadership rather than on broader operations and compliance issues, which is still the norm in many systems" (Burling & Fenton, 2013, para. 6). After spending more than a year working with principals and select district-level leaders to provide professional development on evaluation improvement, the implementation team made the following recommendations as a result of both formal and informal conversations, data collected from the interventions, conclusions drawn by the implementation team, and our own experiences:

- Keep evaluation at the forefront of strategic planning

- Include stakeholders such as the superintendent, the director of human resources, the director of curriculum and instruction, other district-level leaders, and principals in the decisions about evaluation support
- Model best practice by ensuring support includes a variety of activities, opportunities for collaboration, and methods for collecting feedback and reflections
- Allocate time for collaborative discussions, walkthroughs, and standards work throughout the school year
- For participation in the OCT Pilot, develop a specific timeline with benchmarks
- Provide professional development on coaching, feedback, and difficult conversations

Because there is little formal, consistent training on evaluation in North Carolina, and evaluation is one of the elements of the principalship that can provide the impetus for improvement in instructional quality, teacher quality, and ultimately student learning outcomes, evaluation should be a central element of strategic planning for principal support. Educational trends come and go, but evaluation is one constant in all schools. Having the ability to accurately rate the quality of instruction and articulate to a teacher where she is currently performing as well as what she can do to improve is a skill that all site-based administrators need, regardless of the type of school they serve or the state in which they live. Furthermore, a variety of stakeholders need to be involved in the planning of evaluation support. The superintendent helps provide the strategic vision and goals for the district and its leaders. Each district-level leader comes to the table with a variety of experiences and expertise. Each knows the strengths and areas for

improvement they have as well as those the site-based administrators have based on their experiences with them. Developing the district plan to support evaluation should come from the points-of-view and experiences of as many stakeholders involved in evaluation and teaching and learning as possible.

By including opportunities to work with the evaluation standards, engage in collaborative conversations, take part in group observations and discuss both ratings and potential feedback, and provide opportunities for reflection and needs, district-level leaders can help principals grow into more effective evaluators. Principals also understand their strengths and limitations. District-level leaders who create evaluation support structures should be cognizant of principals' needs and ensure that the support and training they provide is aligned both to district-level needs, such as improving target agreement, and to individual areas for improvement. In Concordia, the development of Phase III was a result of both the clearly articulated needs from principals and the data from the OCT Pilot and Fishbowl Intervention.

The OCT Pilot was beneficial to the participants in Concordia. Most participants completed all lessons, and all participants completed all three scoring studies. One of the elements that led to success in Concordia was the implementation of a clear, manageable timeline. Chunking the 34 videos for participants over the course of over four months with seven group-viewing sessions embedded bi-monthly was helpful to participants. When the NCEES Consultant and I asked the participants what they thought about participating in the OCT Pilot, Mrs.Felton shared, "…just looking at that assignment…it looked overwhelming. It was not anything like it appeared to be when you [NCEES Consultant] were here the first time…the way you set it up, it was very easy to manage."

The NCEES Consultant shared with the Concordia participants that many participants across the state had been frustrated because of the lack of district-level support. The clear timeline and expectations, coupled with benchmarks where collaborative discussions took place, improved participation and led to significant improvement on the three measures of the pilot.

Although Concordia participants all agreed that viewing the OCT video lessons together and then discussing the standard, evidences, and wording were the most powerful part of the Phase I intervention, larger districts may have to plan by zone or job-alike group. One of the benefits to leading in a small district is the ability to work with all site-based administrators at one time to develop a shared vision of what instruction should look like and improve evaluation practices through working together to improve collective competence holistically. Mr. Hines also remarked that Concordia's size gave us the opportunity to work in a small collaborative group where all principals had an opportunity to develop collective competence. He commented, "…not everyone's as small as we are. They're not going to be able to pull their entire group like we've been able to do. That's one of the benefits of the video." In a larger district where job-alikes or area-specific principals' meetings are the norm, the OCT Pilot model could be helpful in ensuring alignment of evaluation practices across a district. Thomas Guskey's research (2000) suggests that a clear emphasis on high-quality, ongoing professional development is crucial to changing organizational patterns and norms. Professional development experiences that have these characteristics can lead to improved student outcomes. Sustained support and professional development around evaluation is imperative to calibration and target accuracy. Because North Carolina's teacher

evaluation instrument is so robust, evaluators need ongoing, facilitated support with the language of the standards, what the descriptors look like in a classroom, and how to distinguish between the ratings.

Implications for Policy Makers and the NCDPI

Three consecutive years of data indicate that ratings North Carolina evaluators assign to teachers exemplify the widget effect. Most teachers are rated as good or great, and there is no correlation between student growth and teacher evaluation. More high stakes decisions are being linked to teacher evaluations than ever before, making it more imperative that policymakers address the need for a more comprehensive approach to equipping evaluators to provide teachers with fair, meaningful, and accurate feedback through observation and evaluation (Mullins & Simmons, 2014). Teachers deserve formative feedback following observations that provides them with the information and support they need to improve their practice. However, when student learning outcomes do not correlate with teacher ratings, a review of policy must be considered.

Currently, the three North Carolina's State Board policies around evaluation address only the evaluation process. Without policy designed to ensure teacher evaluators are qualified to provide high-quality feedback, the lack of correlation between student growth and teacher ratings may not improve. Participation in the OCT Pilot and working with evaluators to improve target agreement and reduce both scoring bias and rater discrepancy has revealed an explicit need for policy revisions that should include evaluation certification and ongoing calibration professional development. With the reputation of schools and districts on the line, now, more than ever, "skilled evaluators are essential to a valid and reliable educator evaluation process, and an intentional

investment of time and money is required to cultivate evaluators that can coach teachers to improved practice" (Mullins & Simmons, 2014, p. 1). The following are suggestions policymakers must consider to ensure evaluations are accurate:

- Develop a revised and abbreviated version of the tool for observations
- Create a statewide evaluation certification process so that North Carolina educators may become certified evaluators
- Revise the Observation Calibration Training
- Develop a facilitator's guide for the Observation Calibration Training

Abbreviated Version of the Rubric for Evaluating North Carolina Teachers

With 147 elements, 133 of which are observable, the probability that evaluators will be skilled on all elements is unlikely. Even on the abbreviated evaluation cycle, evaluators must provide feedback on 83 elements, just in standards 1 and 4. The NCEES Consultant mentioned to the Concordia OCT Pilot participants during the Fishbowl Intervention that she intended to develop an observation tool that would be used just for classroom observation, "that gets rid of all that stuff on the rubric that we're not looking for when we go in." She indicated that this tool would require a change in policy. Clarifying fewer, more specific "look-fors" during classroom observations is one way to improve rater accuracy and agreement. Based on the comparative data from SS2 to SS3 on the four elements reviewed during the Fishbowl, a deeper review of fewer elements may lead to more accurate evaluations.

Implementation of an Evaluation Certification Process

Illinois and Ohio currently require evaluators to complete certification trainings. Although their approaches are different, the goal is the same: improve evaluation

practice in terms of calibration. Furthermore, the Ohio model incorporates an additional component of the training, supporting evaluators in providing better feedback to teachers to improve the quality of instruction (Mullins & Simmons, 2014). Both of these states understand that the state does have some direct influence over principals. Those principals have a direct influence over teachers who have a direct influence on student learning outcomes. This adaptation of "The Ripple Effect" provides the impetus for potential needed changes in policy around evaluation licensure and certification.

Policy related to teacher evaluation in North Carolina is limited to three policies all related to the process of evaluation. There is no policy regarding the quality of evaluation at this time. According to McClellan, Atkinson, & Danielson (2012) evaluators must have an understanding of the application of the evaluation rubric, and the rubric itself must be reviewed frequently in order to provide feedback that corrects misunderstandings. Developing policy around the quality of evaluation and ensuring that all North Carolina evaluators receive specific high-quality professional development to improve rater accuracy will improve evaluation practices across the state.

Currently in North Carolina, the only individuals licensed to evaluate teachers are those who have completed a principal preparation program and hold a school administrator principal license. Principal preparation programs differ immensely, and the state does not provide a specific curriculum or guidelines on how to ensure that evaluation training is effective. The quality of training that aspiring administrators receive is contingent on the program. Furthermore, there are many paths to district and school leadership, and many central office leaders do not hold principal licenses but do hold curriculum and instruction certifications or even certifications in educational

leadership. Unfortunately, in North Carolina, these proven leaders are not licensed to evaluate in the state. An evaluation certification process would position practicing site-based and district-level administrators, aspiring educational leaders, and teachers alike to seek evaluation certification in order to improve their practice, gain a deeper understanding of the evaluation standards, and improve their ability to rate accurately without bias.

There are several options for how North Carolina policy makers could develop this certification. First, based on our experience with the OCT Pilot, a fully-online, self-paced model may be the easiest method, but it most likely would not be the most effective. Participants in Concordia expressed that collective thought was powerful and that discussing interpretations, wording, and examples of classroom practice was crucial to their deeper understanding as well as developing collective competence.

Two additional options for North Carolina policymakers to consider include developing a facilitated model that could be implemented at the district level or develop a statewide cohort model facilitated by state evaluation experts.

Locally-facilitated certification model. This train-the-trainer model would be one way to use the OCT model that the state piloted and add specific protocols, questioning strategies, and suggested guidelines to support implementation. District-level facilitators would be leaders who successfully completed the certification training with an evaluation expert from NCDPI. Because the pre-test (SS1)/post-test (SS2) model already exists within the OCT, certification could be determined by the results achieved on the post-test (SS2). Furthermore, participants who did not achieve the minimum score for certification could complete additional modules and then complete an additional

assessment (SS3) as an alternative method of certification. One benefit to this model would be that leaders in the same district would develop the collective competence by engaging in conversations and collaborative training that would support the alignment of evaluation practices throughout the district.

Feedback and coaching component. After participation in a year-long improvement project to support evaluators, one of the most important realizations was a need for principals to have the opportunity to experience ongoing training on providing feedback and coaching. Knowing the tool is important, but being able to provide high-quality, accurate feedback in a manner that is appropriate is another. CPS participants expressed a desire to add elements of the OCT Pilot that included coaching and providing feedback.

Expert-facilitated certification model. Another option is for the state to develop a cohort model much like *Distinguished Leadership in Practice* (DLP), a leadership cohort developed by the North Carolina Principals and Assistant Principals Association, which provides high-quality professional development to leaders across the state in order to improve their practice. The benefit of this model is the consistency in the training that would be offered by the same group of facilitators, thereby ensuring consistency and equity. Furthermore, participants would hear viewpoints and experiences from all over the state, providing more interaction and dialogue among a more diverse group of stakeholders with a variety of experiences. This state-developed cohort model may be perceived as more formal and could be perceived as more meaningful in terms of the prestige of certification. One additional option for this model is to develop the cohorts by region to minimize travel and expense for districts and schools. A regional model could

be supported by the NCDPI NCEES Consultant in conjunction with the regional NCDPI professional development leaders who are assigned to specific state regions. Trainings could be held at the state RESAs and would still provide participants the opportunity to gain insights from a variety of different stakeholders with diverse experiences.

Revision of the Observation Calibration Training

Both the platform and the implementation of the OCT need to be revised prior to implementing the training tool with a state-wide audience. Both Appalachian County Schools and Concordia Public Schools provided specific feedback about improvements in the platform for the OCT that Bloomboard created. These improvements included providing access to view the element(s) under review on the same screen as the video, providing higher-quality instructional materials and lesson plans, and higher-quality video with better audio. Although Casabianca, McCaffrey, Gitomer, Bell, Hamre, & Pianta, (2013) observed that the platform (video or live classroom) for viewing instruction made no difference in rater reliability, the participants in the OCT Pilot would argue that poor video and audio quality did make rating a challenge.

Aside from the platform itself, if NCDPI does implement an evaluator certification process, the OCT should include opportunities for collaboration and collective viewing – whether that be through a cohort model or providing clear, explicit instructions for facilitators about how to implement the OCT effectively.

Development of a Facilitator's Guide. If the OCT is destined to become a part of evaluation support in North Carolina, whether through an evaluation certification process or simply a tool to improve evaluation practices more informally, the NCDPI should develop a facilitator's guide to support effective implementation. The facilitated

model of implementation in Concordia led to improved target agreement, and a reduction in scoring bias and rater discrepancy. A facilitator's guide would provide district- or state-level leaders with specific protocols, tools, and questions to ensure that participants from around the state are provided with a comparable experience with the OCT.

Reflections

Data informed each step of this improvement science project. The implementation and design teams reviewed, analyzed, and made specific decisions for each phase based on the data from previous phases. Practitioner research relies on data to drive each step of improvement. Working with the North Carolina Department of Public Instruction and having access to measurement tools that provided quality data ensured that each decision was made in response to data from the previous phase. Improvement science calls for practitioners to develop plans for implementation but also allows for the purposeful abandonment of the original project design in response to the data and needs of the institution or stakeholders. This tenet of improvement science not only guided this project but was exhibited throughout the project as layers of data revealed information that led to modifications in the original design plan and future iterations to come.

Research in the field of evaluation improvement by Joe, Kosa, Tierney, & Tocci (2014) indicates that evaluators should be engaged in observation practice activities a minimum of three times per year. It is critical to provide evaluators with an opportunity to discuss their interpretations and improve their understanding of the Rubric for Evaluating North Carolina Teachers through job-embedded professional development. Through these experiences, evaluators can create a common language and improve the quality of evaluation in the organization. Best practice also includes comparing

evaluators' evidence and ratings to those provided by master scorers and analyzing video examples of teaching to develop rating competence (McClellan et al., 2012). The OCT Pilot provided five months of ongoing opportunities for Concordia evaluators to participate in observation practices that included these improvement opportunities. Furthermore, participating in job-embedded observations of classroom practice is an important part of improving evaluation practice, as evidenced by the results of this project. Concordia's CWT process and the Phase II Fishbowl Intervention provided evaluators with the opportunity to improve their evaluation skills by visiting real classrooms and taking time to discuss those experiences as they pertained to the elements and concepts under review. Furthermore, evaluators need opportunities to demonstrate their proficiency periodically to improve rater reliability (Bell, Qi, Croft, Leusner, McCaffrey Gitomer, & Pianta, 2014). Concordia evaluators had multiple opportunities to demonstrate their proficiency – the OCT Mini-Pilot scoring study, SS1, SS2, and SS3. The design of the Concordia evaluation improvement project incorporated the criteria set forth by researchers in the field in an effort to yield the most significant improvement.

Future interventions will continue to take into account the research conducted by experts in the field but will also incorporate lessons learned from this project. These lessons include:

- Ensuring evaluation is a focus at monthly curriculum principals' meetings by either completing collaborative walkthroughs followed by small group discussions using our district-wide CWT document or a specific set of elements from the NCEES Rubric
- Including assistant principals in evaluation support opportunities

- Using both the *Evidences for Professional Teacher Standards 1-5* and the *NCEES: Questions for Post-Observation Conference and Summative Evaluation* documents for talking points during collaborative evaluation conversations

We will also include a specific emphasis on support with coaching conversations and on providing high-quality instructional feedback to teachers. Evaluators in Concordia articulated a need for more focused support with the post-observation conference. If leaders are committed to working with principals to improve the quality of evaluation, teacher quality will improve. With an improvement in teacher quality, student learning outcomes will also improve. As we continue to work to improve our evaluation practices, our goal is to learn from each intervention in order to meet the individual and collective needs of our principals so that, in the indirect manner described in our adaptation of *The Ripple Effect*, we can impact classroom instruction and ultimately student learning.

REFERENCES

ASA statement on using value-added models for educational assessment. (2014, April 8). Retrieved from https://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf

Batton, D., Britt, C., DeNeal, J., & Hales, L. (2012). *NC teacher evaluations & teacher effectiveness: Exploring the relationship between value-added data and teacher evaluations* (*Project 6.4*). Retrieved from http://www.ncpublicschools.org/docs/intern-research/reports/teachereval.pdf

Bell, C.A., Gitomer, D.A., McCaffrey, D., Hamre, B., Pianta, R., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment,* 17(2-3), 62-87.

Bell, Courtney A., Yi Qi, Andrew J. Croft, Dawn Leusner, Daniel F. McCaffrey, Drew H. Gitomer, and Robert C. Pianta (2014). Improving Observational Score Quality: Challenges in Observer Thinking. In T. J. Kane, K. A. Kerr & R. C. Pianta (Eds.) *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project*, San Francisco, CA: Jossey-Bass.

Bill & Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three year study*. Retrieved from http://metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf

Bolman, L., & Deal, T. (2008). *Reframing organizations: Artistry, choice, and leadership* (4th Ed.). San Francisco, CA: Jossey-Bass.

Brookover, W. B., & Lezotte, L. (1982). *Creating effective schools*. Holmes Beach, FL: Learning Publications.

Burling, K., & Fenton, B. (2013, September 27). Connecting principal evaluation and professional development. Retrieved from http://researchnetwork.pearson.com/educator-effectiveness/connecting-principal-evaluation-and-professional-development

Bryk A. S., Gomez L. M., & Grunow A. (2011), *Getting ideas into action: Building Networked Improvement Communities in education*. [Essay]. Carnegie Foundation for the Advancement of Teaching, Stanford, CA. Retrieved from http://www.carnegiefoundation.org/

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychology Measurement*, *73*, 757.

Clifford, M., Behrstock-Sherratt, E., & Fetters, J. (2012). *The ripple effect: A synthesis of research on principal influence to inform performance evaluation design. A quality school leadership issue brief*. American Institutes for Research. Retrieved from http://files.eric.ed.gov/fulltext/ED530748.pdf

Darling-Hammond, L. (2006). *Powerful teacher education: Lessons from exemplary programs*. San Francisco, CA: Jossey-Bass.

Darling-Hammond, L. (2015). Can value added add value to teacher evaluation? *Educational Researcher*, *44*(2), 132–137.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E. H., & Rothstein, J. (2011). *Getting teacher evaluation right: A background paper for policy makers*. Retrieved from http://iaase.org/Documents/Ctrl_Hyperlink/Session_30c_GettingTeacherEvaluationRight_uid9102012952462.pdf

Desimone, L. M. (2009). Improving impact studies of teachers' professional development toward better conceptualizations and measures. *Educational Researcher, 38*(3), 181-199.

Empirical Education. (2015a). *Concordia Public Schools scoring study 1 report*. [Unpublished raw data].

Empirical Education. (2015b). *Concordia Public Schools scoring study comparisons*. [Unpublished raw data].

Evidences of professional teacher standards 1-5. (2014, October 10). Retrieved from: http://ncees.ncdpi.wikispaces.net/file/view/Evidences-READYprincipalsMtg_UPDATED.pdf/525990024/Evidences-READYprincipalsMtg_UPDATED.pdf

Gandha, T., & Baxter, A. (2015). *Toward trustworthy and transformative classroom observations: Progress, challenges, and lessons in SREB states*. Atlanta, GA: Southern Regional Education Board, 2015.

Goldhaber, D. (2008). Teachers matter, but effective teacher policies are elusive. In *Handbook of Research in Education Finance and Policy*. New York, NY: Routledge.

Guskey, T. R. (2000). *Evaluating professional development.* Thousand Oaks, CA: Corwin.

Guskey, T. R. (2009). Closing the knowledge gap on effective professional development. *Educational Horizons, 87*(7); 224-233.

Guskey, T. R. & Yoon, K. S. (2009). What works in professional development? *Phi Delta Kappan 90*(7), 495-500.

Higgins, L., H., Marks, J., Barrett, N., Henry, G.T., Guthrie, S., & Comperatore, A. (September, 2013). *Measures of student growth in the educator evaluation system: A formative report.* Chapel Hill, NC: Education Policy Initiative at Carolina.

Ho, A., & Kane, T. (2013). *The reliability of classroom observations by school personnel*. Retrieved from http://www.metproject.org/downloads /MET_Reliability%20of%20Classroom%20Observations_Research%20Paper.pdf

Joe, J., Kosa, J., Tierney, J., & Tocci, C. (2014). *Observer calibration: A tool for maintaining accurate and reliable classroom observations*. San Francisco, CA: Teachscape.

King, A. (2008). In vivo coding. In L. Given (Ed.), *The SAGE encyclopedia of qualitative research methods.* (pp. 473-474). Thousand Oaks, CA: SAGE Publications, Inc. http://dx.doi.org/10.4135/9781412963909.n240

Kruse, S. D. & Louis, K. S. (2009). *Building strong school cultures: A guide to leading change*. Thousand Oaks, CA: Corwin Press.

Leadership matters: What the research says about the importance of principal leadership. (2013). Restin, VA: NASSP.

Leithwood, K., Seashore Louis, K., Anderson, S. & Wahlstrom, K. (2004). How leadership influences student learning. New York: Wallace Foundation. Retrieved from http://www.wallacefoundation.org/

Marzano, R. J., Waters, T., & McNulty, B. A. (2005). *School leadership that works: From research to results*. Alexandria, VA: ASCD.

McCaffrey, J. R., Lockwood, D. F., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value added models for teacher accountability* [Monograph]. Santa Monica, CA: RAND Corporation. Retrieved from http://www.rand.org/pubs/monographs/2004/RAND_MG158.pdf

McClellan, C., Atkinson, M., & Danielson, C. (2012). *Teacher evaluator training and certification: Lessons learned from the Measures of Effective Teaching project*. San Francisco, CA: Teachscape.

McDonald, J.H. (2014). *Handbook of Biological Statistics.* (3rd Ed.). Baltimore, MD: Sparky House Publishing.

McGuinn, P. (2012). *The state of teacher evaluation reform: State education agency capacity and the implementation of new teacher-evaluation systems*. Retrieved from http://www.americanprogress.org/wp-content/uploads/2012/11 /McGuinn_TheStateofEvaluation-1.pdf

Merriam, S. B. (2001). Andragogy and self-directed learning: Pillars of adult learning theory. *New Directions for Adult & Continuing Education, 89*: 3-14.

Mullins, H. P, & Simmons, K. B. (2014). *Improving the quality of educator evaluation in North Carolina through policy analysis and revision*. Retrieved from

http://ncees.ncdpi.wikispaces.net/file/detail/Mullins-Simmons+Policy+Brief+Final+(2)+(1).docx

Our ideas: The six core principles of improvement. (2016). Retrieved from http://www.carnegiefoundation.org/our-ideas/

Peterson, K. & Cosner, S. (2005). Teaching your principal: Top tips for the professional development of the school's chief. *Journal of Staff Development, 26*(2): 28-30.

Popham, W.J. (2010). *Everything school leaders need to know about assessment.* Thousand Oaks, CA: Corwin.

Public School Forum of North Carolina (2013). *Local school finance study.* Retrieved from http://www.ncforum.org/wp-content/uploads/2013/05/FinanceStudy_Dec2013_final.pdf

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2000). Teachers, schools, and academic achievement (Working Paper W6691). Cambridge, MA: National Bureau of Economic Research.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record, 104*, 1525-1567.

Roy, P., & Hord, S. (2003). *Moving NSDC's staff development standards into practice: Innovation configurations* (Vol. I). Oxford, OH: NSDC.

Sandelowski, M., Voils, C. I., & Knafl, G. (2009). On quantitizing. *Journal of Mixed Methods Research, 3*(3), 208-222.

Sanders, W. L., Wright, S. P, Rivers, J. C., & Leandro, J. G. (2009*). A response to criticisms of SAS EVAAS*. [White paper]. Retrieved from

http://www.sas.com/resources/asset/Response_to_Criticisms_of_SAS_EVAAS_11-13-09.pdf

Simmons, K. B. (2014, June 30). NCDPI Observation Calibration Training [Webinar]. Retrieved from http://ncees.ncdpi.wikispaces.net/Archived+Webinars+2014-2015

Simmons, K. B. (2014, September). *Deep dive for the observation calibration training*. [Unpublished raw data].

Simmons, K. B., & Mullins, H. P. (2013). Adaption of "The Ripple Effect." Retrieved from http://edlstudio.wikispaces.com/file/view/ripple.png/527174698/ripple.png

Strong, M., Gargani, J., & Hacifazlioğlu, Ö. (2011). Do we know a successful teacher when we see one? Experiments in the identification of effective teachers. *Journal of Teacher Education, 62*(4); 1-16.

Taylor, E. S., & Tyler, J. H. (2012). Can teacher evaluation improve teaching? *Education Next, 12*. Retrieved from http://educationnext.org/

Tomberlin, T. (2014). READY principals. *NCEES*. Retrieved from http://ncees.ncdpi.wikispaces.net/READY+Principals+Spring+2014

U.S. Department of Education. (2004). *New No Child Left Behind flexibility: Highly qualified teachers*. Retrieved from http://www2.ed.gov/nclb/methods/teachers/hqtflexibility.html

Value-added measures in teacher evaluation. (2014, November 7). Retrieved January 18, 2015, from http://www.nassp.org/Content.aspx?topic=Value_Added_Measures_in_Teacher_Evaluation

von Hippel, E. (2005). *Demoralizing innovation.* Cambridge, MA: The MIT Press.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. Retrieved from http://widgeteffect.org/downloads /TheWidgetEffect.pdf

West-Ed. (2012). *Options for incorporating student academic growth as one measure of the effectiveness of teachers in tested grades and subjects: A report to the North Carolina Department of Public Instruction*. Retrieved from http://www.ncpublicschools.org/docs/effectiveness-model/evaas/selection/options-for-incorporating.pdf

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teachers and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 57-67.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Issues & Answers Report, REL 2007, no. 033).Washington, D.C.: U.S. Department of Education, Institute of Education Sciences. Retrieved from: http://ies.ed.gov/ncee/edlabs

APPENDICES

Appendix A. *North Carolina Teacher Evaluation Process Manual* - http://bit.ly/nceesmanual

Appendix B. 2013-2014 Concordia Classroom Walkthrough Tool - https://docs.google.com/a/nccs.k12.nc.us/forms/d/1qgMnTGFwE4vM5gBOYPlFIBtUsqqV5KELYJ9cSKEqd90/viewform?fbzx=-589360558307605744

Appendix C. NCEES Wikispace - www.ncees.ncdpi.wikispaces.net

Appendix D. *TCP-C-004 –Policy Establishing the Teacher Performance Appraisal Process* - http://sbepolicy.dpi.state.nc.us/policies/TCP-C-004.asp?pri=02&cat=C&pol=004

Appendix E. *TCP-C-006 – Policy on Standards and Criteria for Evaluation of Professional School Employees* - http://sbepolicy.dpi.state.nc.us/policies/TCP-C-006.asp?pri=02&cat=C&pol=006&acr=TCP

Appendix F. *TCS-C-021 – Policy on Educator Value-Added Assessment System (EVAAS) Teacher Module* - http://sbepolicy.dpi.state.nc.us/policies/TCS-C-021.asp?pri=04&cat=C&pol=021&acr=TCS

Appendix G. National School Reform Faculty *Chalk Talk* Protocol - http://www.nsrfharmony.org/system/files/protocols/chalk_talk_0.pdf

Appendix H. National School Reform Faculty *Critical Friends* Protocol - http://www.nsrfharmony.org/system/files/protocols/cfg_purpose_work_0.pdf

Appendix I. Concordia Protocol for Facilitated Lessons - https://docs.google.com/document/d/1kTwBpW9H_i_DqjUdEMqgao9Q_U-j6p3qJw-xnSfLHcA/edit

Appendix J: Sample OCT Element Handout - https://docs.google.com/document/d/1nokkuGC7zg_cV8dmMdBr8bUkeAnKs3Kv0NAlE0bKUek/edit

Appendix K: Concordia Raw Data from OCT Pilot Scoring Studies - https://docs.google.com/spreadsheets/d/1lpEEii9VGGGBYq7OvYQ3k7j-Ha0A1oNV6eM-IyFAE6E/edit#gid=0

Appendix L: Guskey's Five Critical Levels of Professional Development Evaluation - https://connectingcantycommunities.wikispaces.com/file/view/Guskey+5+levels.pdf

Appendix M: "Capture Your Thoughts" Handout - https://docs.google.com/document/d/15mUTAq48uAL7a1jwW8Fs3BvYoKCqlIJrFeMLmGxySRQ/edit

Appendix N: Selection of Standards Anecdotal Notes - https://docs.google.com/spreadsheets/d/1KvaGAXBIC_mC9MUgMTDIzgDrLETm0VpvSVe31xxFRIk/edit#gid=0

Appendix O: 2015-2016 Concordia Classroom Walkthrough Tool - https://docs.google.com/forms/d/1bNK0IoR41-Kzud6cpJX-cYjuLAznciDhStwfg7H5M_M/viewform

Appendix P: "NCEES: Questions for Post-Observation Conference and Summative - Evaluation" http://ncees.ncdpi.wikispaces.net/file/view/NCEES-questions.pdf

Appendix Q: "Evidences for Professional Teacher Standards 1-5" http://bit.ly/1RyrQwS